

Computational surprisal is not enough: Discourse prominence modulates human eye-movement patterns but not GPT expectations in Korean

Nayoung Kwon Gyu-Ho Shin Seongmin Mun
Kyungpook National University University of Illinois Chicago Kyungpook National University
nayoungkw@knu.ac.kr ghshin@uic.edu seongminmun@knu.ac.kr

Abstract

Large Language Model (LLM) surprisal is increasingly used as a proxy for human sentence processing, yet whether model expectations reflect the cognitive factors governing real-time human behavior remains uncertain. We examine Korean subject–verb honorific agreement in subject- and object-control constructions, where human eye-tracking data reveal a prominence-driven asymmetry. Evaluating three Korean-capable models, we find that surprisal was sensitive to honorific-feature mismatches, yielding attraction-like effects, but was not modulated by distractor prominence. Although surprisal selectively improved reading-time predictions, this reflected item-level predictive overlap rather than prominence sensitivity. GPT models thus capture surface regularities but do not reproduce the construction-sensitive, prominence-driven asymmetry of human dependency resolution, suggesting that LLMs diverge from humans precisely where processing depends on working-memory constraints rather than distributional expectation.

1 Introduction

Computational methods have become integral to language sciences, offering a new window to modeling human language behavior and investigating the cognitive mechanisms underlying language processing (Contreras Kallens et al., 2023; Frank and Goodman, 2026; Hu et al., 2020; Jones and Bergen, 2024; Marvin and Linzen, 2018; Warstadt and Bowman, 2020; Wilcox et al., 2023; Yoo et al., 2025; cf. Bavaresco et al., 2025). Much of this research is grounded in surprisal theory (Hale, 2001; Levy, 2008), which posits that processing difficulty is strictly proportional to conditional unexpectedness. While surprisal models successfully account for various patterns of human reading (Smith and Levy, 2013; Wilcox et al., 2023), a fundamental question remains: do complex sentence-processing

phenomena—such as asymmetry in agreement attraction—emerge naturally from expectation-based processing over distributional input, or do they depend substantially on algorithmic constraints of the human memory architecture (Lewis and Vasishth, 2005; Vasishth et al., 2019)? Evaluating language model surprisal against human behavioral data offers a powerful theoretical diagnostic for disentangling expectation-driven processing costs from effects arising from memory constraints. In the present study, we pursue this inquiry by examining how discourse prominence modulates agreement attraction in the processing of long-distance Korean subject–verb honorific dependencies.

1.1 Attraction phenomenon

Human sentence comprehension often requires establishing dependencies between linguistic elements that are separated by intervening items; a large body of psycholinguistic research suggests that such dependencies are resolved through a memory retrieval process (Lewis and Vasishth, 2005; Van Dyke and McElree, 2006).

Specifically, cue-based retrieval models propose that the parser retrieves the licit target item by matching cues associated with the current retrieval probe (e.g., number or gender features) against representations stored in a content-addressable memory. During this process, structurally inappropriate candidates may also be retrieved due to partial matches with the retrieval cues, yielding interference effects (McElree, 2000; Lewis and Vasishth, 2005; Vasishth et al., 2019). When such misretrieval occurs, it often manifests as a reduction in processing difficulty in ungrammatical sentences where the licit target mismatches the retrieval probe, a phenomenon known as facilitatory interference or agreement attraction (Pearlmutter et al., 1999; Wagers et al., 2009). Such effects are well-documented across languages (Franck et al., 2015; Lago et al., 2015; Dillon et al., 2013), and Ko-

rean is no exception, as distractors whose honorific features match the retrieval cues induce facilitatory interference in otherwise ungrammatical sentences (Kwon and Sturt, 2016, 2019).

1.2 Discourse prominence

Human language processing is constrained by working memory capacity, and consequently, memory representations may be activated to varying degrees during retrieval. Distractor interference effects may therefore not be uniform across sentence configurations, reflecting this variability in memory activation (Jäger et al., 2017). For example, a distractor closer to the retrieval position elicits stronger interference than one farther away (Sturt and Kwon, 2024). This finding is consistent with the Adaptive Control of Thought-Rational framework (ACT-R; Lewis and Vasishth, 2005), which attributes this locality effect to the decay of memory activation over time.

However, distance is not the only determinant of memory activation. In an extension of ACT-R, Engelmann et al. (2019) argue that discourse or sentential prominence plays a critical role in retrieval probability. Specifically, an NP occupying a prominent position such as a sentential subject or discourse topic (Ariel, 1990; Chafe, 1976; Givón, 1983) is maintained with a high level of base memory activation and therefore retrieved with higher probability than an NP in a less prominent position. Accordingly, a distractor in a prominent position is predicted to elicit stronger attraction effects than one in a less prominent position, a prediction supported by the meta-analysis of attraction studies reported in Jäger et al. (2017) and by simulations of the revised ACT-R model against the studies examined therein (Engelmann et al., 2019). Evidence from Korean subject honorification further supports this account: attraction effects were stronger with the subject distractor NP than with an object distractor NP, suggesting that sentential prominence is a weighted retrieval cue (Kwon and Sturt, 2019).

2 Related Work

Attraction-like effects have been observed in LLMs. Ryu and Lewis (2021) demonstrated that surprisal values derived from GPT-2 predict facilitatory interference effects in ungrammatical sentences, where a grammatically unlicensed NP matching the verb in number leads to reduced surprisal. Bacon and Regier (2019) similarly found that, although LLMs

achieve high classification accuracy across 26 languages, model performance declines systematically in the presence of agreement attractors. Lasri et al. (2022) further showed that Transformer-based models do not generalize in a fully lexically independent manner, particularly when processing sentences containing attractors. Chaves and Richter (2021) also observed that, although LLMs encode rich syntactic representations, they frequently resort to shallow heuristics rather than genuine structural computation. In Korean subject–verb honorific agreement, Kwon and Mun (2025) corroborated this pattern: while KoGPT-2 outperformed KoBERT on structurally complex constructions, both models remained susceptible to interference from honorific attractors, resulting in reduced grammaticality classification accuracy.

These findings indicate that attraction-like effects emerge systematically across LLM architectures and agreement domains. Nevertheless, it remains an open question whether these effects exhibit the prominence sensitivity that characterizes human attraction, or arise from fundamentally different sources rooted in statistical optimization. The present study aims to address this gap by examining the effect of discourse prominence on LLM surprisal patterns.

While the role of discourse prominence in modulating attraction effects has received growing empirical support in human sentence processing, its counterpart in LLMs remains largely unexplored. This gap is notable because both LLMs and human comprehenders exhibit both locality and attraction effects, with performance declining with longer dependencies and in the presence of distractors (Bacon and Regier, 2019; Chaves and Richter, 2021; Ryu and Lewis, 2021; Lasri et al., 2022; see also Shi et al., 2023, for similar findings in mathematical contexts), a pattern that extends to Korean subject–verb honorific agreement (Kwon and Mun, 2025).

A particularly relevant study is Cuneo et al. (2025), who examined LLMs’ sensitivity to information structure. Manipulating prominence through explicit capitalization (e.g., write HER vs. write her), they found that GPT-4’s acceptability ratings varied systematically, mirroring human ratings. Capitalization, however, is itself an overt surface cue that a model can track as a token-level regularity without encoding prominence structurally. The paradigm thus cannot separate sensitivity to a surface marker from the role-based prominence

(1) Example sentences (Kwon and Sturt, 2019)

(a) Subject-control construction (Experiment 1)

‘The teacher_i/Swucin_i emphasized to the chairperson_k/Minhyeong_k that PRO_i would wait ...’

선생님/수진-이 위원장/민형-에게 ... PRO_i 기다리-시-겠다고 강조하셨다.
teacher/Swucin-NOM chairperson/Minhyeong-DAT wait-HON-SUBJ.CTRL.COMP emphasized

(b) Object-control construction (Experiment 2)

‘The teacher_i/Swucin_i emphasized to the chairperson_k/Minhyeong_k that PRO_k should wait ...’

선생님/수진-이 위원장/민형-에게 ... PRO_k 기다리-시-라고 강조하셨다.
teacher/Swucin-NOM chairperson/Minhyeong-DAT wait-HON-OBJ.CTRL.COMP emphasized

that drives human retrieval, leaving open whether prominence differences with no surface exponent (e.g., those inferable only from grammatical role, such as the subject–object asymmetry) modulate model behavior as they do in humans.

Since LLMs lack human-like working memory constraints, there is no architectural pressure for them to differentially weight prominent NPs during retrieval. This leaves open the question of whether they develop the same asymmetric prominence sensitivity observed in human processors.

3 Current Study

Building on Kwon and Sturt (2019), the present study evaluates prominence-driven processing in LLMs by employing Korean control constructions (1)¹ as a diagnostic ‘stress test’. These structures provide a robust testbed because the controller of the embedded null subject (PRO) varies by construction type, while the two constructions differ minimally in surface form yet reverse the structural prominence of the distractor. This design allows us to isolate role-based prominence from the surface cues that models are known to exploit (Cuneo et al., 2025). Kwon and Sturt (2019) found that, in subject-control sentences (1a), where the distractor occurs in the object position and thus carries lower prominence, interference from the distractor NP did not emerge in any measure. In contrast, in object-control sentences (1b), where the distractor occurs in the subject position and thus carries higher prominence, the interference effect was strong and emerged in regression-path duration even before the grammaticality effect.

¹Abbreviations: NOM = nominative; DAT = dative; HON = honorific; SUBJ.CTRL = subject control; OBJ.CTRL = object control; COMP = complementizer.

We address the following research question using a focused theoretical diagnostic: *Do GPT language model surprisal patterns exhibit the prominence-driven asymmetry predicted by the revised ACT-R retrieval model (Engelmann et al., 2019), or do attraction-like surprisal effects operate symmetrically regardless of structural prominence?* We predict that, if these models are sensitive to discourse prominence in a manner analogous to human prominence-sensitive retrieval, their surprisal patterns should mirror the asymmetry in human sentence processing (Kwon and Sturt, 2019). This would be indicated by stronger interference in object-control than in subject-control constructions. Conversely, if model behavior is not shaped by such memory-based prominence sensitivity, these models may fail to reproduce this construction-specific asymmetry, instead showing symmetric interference or patterns driven by linear distance rather than structural prominence.

4 Methods

4.1 Materials and computational workflow

The experimental stimuli were adopted from Kwon and Sturt (2019) and comprised subject-control and object-control constructions (1).² Korean marks subject honorification with the verbal suffix *-si*, which agrees with an honorifiable subject. In the control constructions examined here, the embedded predicate carries *-si* along with a control-marking morpheme that determines which matrix argument controls the embedded null subject (PRO). Specifically, subject control is encoded by *-keyss* and object control by *-la*: in subject-control sentences,

²The surprisal-computation code, per-item surprisal values for all three language models, and all analysis scripts are available at <https://osf.io/rs745>.

Table 1: Information about three GPT language models

	KoGPT-2	Ko-GPT-Trinity	mGPT
Parameter #	125M	1.2B	1.3B
Layer #	12	24	24
Head #	12	16	16
Hidden layer dimension	768	2,048	2,048
Feedforward network dimension	3,072	8,192	8,192

PRO is controlled by the matrix subject, whereas in object-control sentences, it is controlled by the dative NP. Because *-si* on the embedded verb must agree with the honorific features of the controller, the two constructions establish the honorific-agreement dependency on structurally different NPs while holding linear order constant. We employed a 2×2 factorial design manipulating the honorific features of the matrix subject (Honorifiable [H] vs. Non-honorifiable [NH]) and the dative NP (H vs. NH), resulting in four conditions: H-H, NH-H, H-NH, and NH-NH.

For modeling, we employed three GPT variants that support Korean: KoGPT-2³, Ko-GPT-Trinity⁴, and mGPT⁵. Table 1 presents detailed specifications of each model. Because these models differ in scale and in whether pretraining was Korean-specific or multilingual, they do not isolate any single factor; we therefore treat them as a generality check, asking whether the same qualitative pattern holds across models rather than comparing them directly. To ensure a zero-shot evaluation of processing patterns, sentences were processed in a ‘closed-loop’ environment without task-specific fine-tuning or prompt engineering.

The workflow began by loading the three models via Hugging Face. Because different architectures often employ unique text-processing methods, we standardized tokenizer behavior to ensure a consistent computational framework across the three models. This involved two primary adjustments. First, we configured the models for left-padding and verified that each model always processes the entire preceding history of a sentence before predicting the subsequent token. Second, we established a consistent method for identifying sentence boundaries (*<start>* and *<end>* tokens), ensuring that the results remained comparable across models.

Next, to maintain the integrity of the linguistic stimuli, each sentence was segmented into *ejels*⁶ via whitespace, followed by subword tokenization. Notably, we did not remove punctuation. Because segmentation relied solely on whitespace, punctuation marks were treated as part of the constituent *ejel* and subsequently handled by the model-specific subword tokenizers. This approach ensured that the surprisal metrics reflected the naturalistic input available to the models, including the structural cues provided by punctuation, which are essential for determining sentence boundaries and clausal finality.

Once tokens were mapped to numerical identifiers and assembled into input sequences, each model generated next-token probability distributions to calculate the surprisal for each token. Surprisal was computed as the negative logarithm of token probabilities conditional upon their preceding context. By using base 2, the resulting surprisal values were expressed in bits, a standard unit of information that allows for an intuitive comparison of information density across models.

Finally, *ejel*-level surprisal was obtained by aggregating the surprisal values of all constituent subwords within a whitespace-delimited segment, capturing the total processing cost associated with each linguistic unit. Crucially, across all four experimental conditions, the subword token counts within the critical predicate and spillover *ejels* were identical within each model’s respective tokenizer. This controlled segmentation ensured that condition-level surprisal variations reflected genuine conditional probability shifts rather than artifacts of subword length or tokenization density.

³<https://huggingface.co/skt/kogpt2-base-v2>

⁴<https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5>

⁵<https://huggingface.co/ai-forever/mGPT>

⁶An *ejel*, comprising one or more morphemes, is a Korean writing unit typically separated by spaces. For example, *na-nun* ‘I-TOP’ in *na-nun khi-ka khu-ta* ‘I-TOP height-NOM tall-DCL’ (‘I am tall.’) is one *ejel* in Korean.

4.2 Statistical analysis

All analyses were conducted in R 4.5.2 (R Core Team, 2025). Mixed-effects models were fitted using lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). Categorical predictors were sum-coded throughout. Random slopes for fixed effects were included where supported by the data and model convergence (Barr et al., 2013).

We conducted two types of surprisal analyses. First, to evaluate whether surprisal at the critical region differed across conditions, we fitted linear mixed-effects models separately for each language model and construction type, with fixed effects for the honorific features of the main-clause subject and the dative NP (H vs. NH) and their interaction. By-item random slopes were included where supported by convergence; the full random-effects structure for every regression is reported in Table E. Second, to compare the magnitude of attraction effects across constructions, we combined the subject- and object-control data and, within each construction, contrasted the attraction condition with its grammatical counterpart (subject-control: NH-H vs. H-NH; object-control: H-NH vs. NH-H). For each language model, we fitted a model with fixed effects for Construction, this attraction contrast, and their interaction (random-effects structure as in Table E). The Construction $\times$ attraction-contrast interaction tested whether the attraction effect differed in magnitude between constructions.

To evaluate whether language-model surprisal predicts human sentence-processing difficulty, surprisal estimates from the three models were incorporated into mixed-effects models of the eye-tracking data reported in Kwon and Sturt (2019). The dependent variables were three standard eye-tracking measures: first-pass reading time, regression-path duration, and total reading time. All measures were extracted from the critical region comprising the predicate and the following word (predicate + 1). For each measure, a baseline model was first fitted with fixed effects of the honorific features of the main-clause subject and the dative NP, as well as their interaction. Surprisal values from each language model were then added as predictors in separate models. To assess whether the inclusion of surprisal significantly improved model fit, baseline and surprisal-augmented models were compared using likelihood-ratio tests implemented with the anova() function. A significant

likelihood-ratio test was interpreted as evidence that surprisal explained additional variance in human reading times beyond the structural predictors alone. These analyses were conducted separately for subject-control and object-control structures for each of the three reading-time measures.

As non-significant interaction contrasts cannot distinguish genuine statistical symmetry from insufficient power, frequentist analyses were supplemented with Bayesian mixed-effects regressions fitted using brms (Bürkner, 2017). Models mirrored the combined LME specifications, utilizing four Markov chains with 10,000 iterations each (2,000 warmup) and weakly informative priors ($\mathcal{N}(0, 2)$) on fixed effects. All models achieved full convergence ($\hat{R} = 1.00$). Hypotheses were evaluated using Savage–Dickey density ratios to compute Bayes factors (BF_{01}) quantifying evidence favoring the null hypothesis (H_0 : absence of construction-dependent modulation) over the alternative (H_1).

5 Results

5.1 Human eye movement patterns

Human data, as reported in Kwon and Sturt (2019), revealed that distractor interference was modulated by construction type in a manner consistent with prominence-driven retrieval (see Figure 1). In subject-control sentences, where the distractor occupies a low-prominence object position, interference was not significant. In contrast, object-control sentences, where the distractor occupies a high-prominence subject position, exhibited robust and early interference. This effect emerged in regression-path durations prior to the effect of the structurally licensed controller and persisted into total reading times. These patterns suggest that the human parser is more strongly influenced by discourse prominence than by structural licensing or linear proximity, aligning with the predictions of the revised ACT-R model (Engelmann et al., 2019).

5.2 GPT surprisal analysis

Figure 2 presents mean surprisal values per condition within construction type by language model, and Table 2 summarizes statistical model outputs (see also Appendices A and C for complete analysis results). In subject-control constructions, the three models exhibited highly similar patterns: surprisal increased monotonically as a function of the number of non-honorific elements in the sentence, with the lowest values observed in the H-H condition

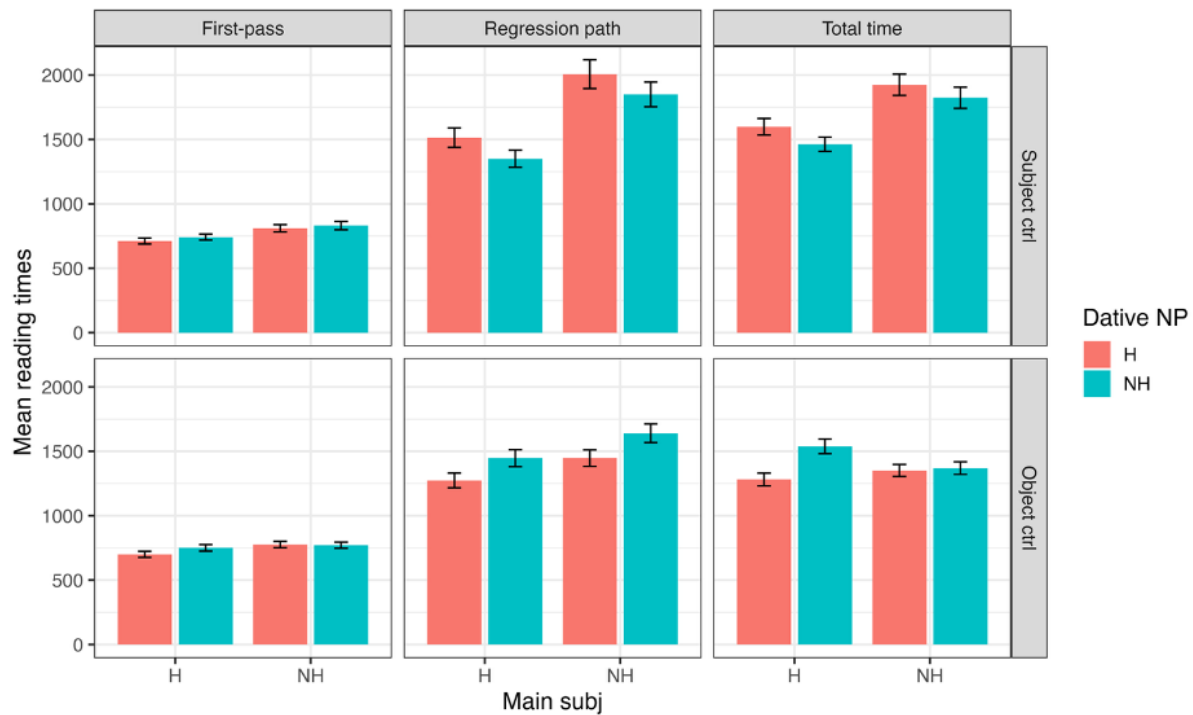


Figure 1: Mean human reading times per condition for subject- and object-control constructions from Experiments 1 and 2 (adapted from Kwon and Sturt, 2019). Error bars represent ± 1 standard error of the mean.

and the highest in the NH-NH condition. Importantly, the pattern did not change as a function of distractor prominence, as indicated by comparable surprisal values in the H-NH and NH-H conditions. Mixed-effects analyses confirmed this pattern. For each model, the main effects of Main Subject and Dative NP were significant, as was the interaction between them. Follow-up pairwise comparisons further showed that all condition contrasts were significant, except for the H-NH versus NH-H contrast, indicating comparable surprisal when only one NP was non-honorifiable.

In object-control constructions, KoGPT-2 and mGPT showed patterns similar to those found in subject-control: a monotonic increase in surprisal

with the number of non-honorific elements, with comparable values in the H-NH and NH-H conditions. Trinity, in contrast, showed an additional pattern beyond the monotonic increase: the H-NH condition elicited significantly higher surprisal than the NH-H condition, in the direction of the grammaticality effect observed in human total reading times. Mixed-effects analyses again revealed significant main effects of Main Subject and Dative NP and a significant interaction for each model. Follow-up pairwise comparisons showed that the H-NH versus NH-H contrast was non-significant for KoGPT-2 and mGPT but significant for Trinity. However, in the combined analysis, the corresponding simple contrast for Trinity in object control

Table 2: Summary of mixed-effects model results for GPT surprisal in subject- and object-control constructions. * $p < .05$, ** $p < .01$, *** $p < .001$.

Construction	Model	Main Subject	Dative NP	Interaction
Subject control	KoGPT-2	***	***	***
	Ko-GPT-Trinity	***	***	***
	mGPT	***	***	***
Object control	KoGPT-2	***	***	**
	Ko-GPT-Trinity	***	***	***
	mGPT	***	***	***

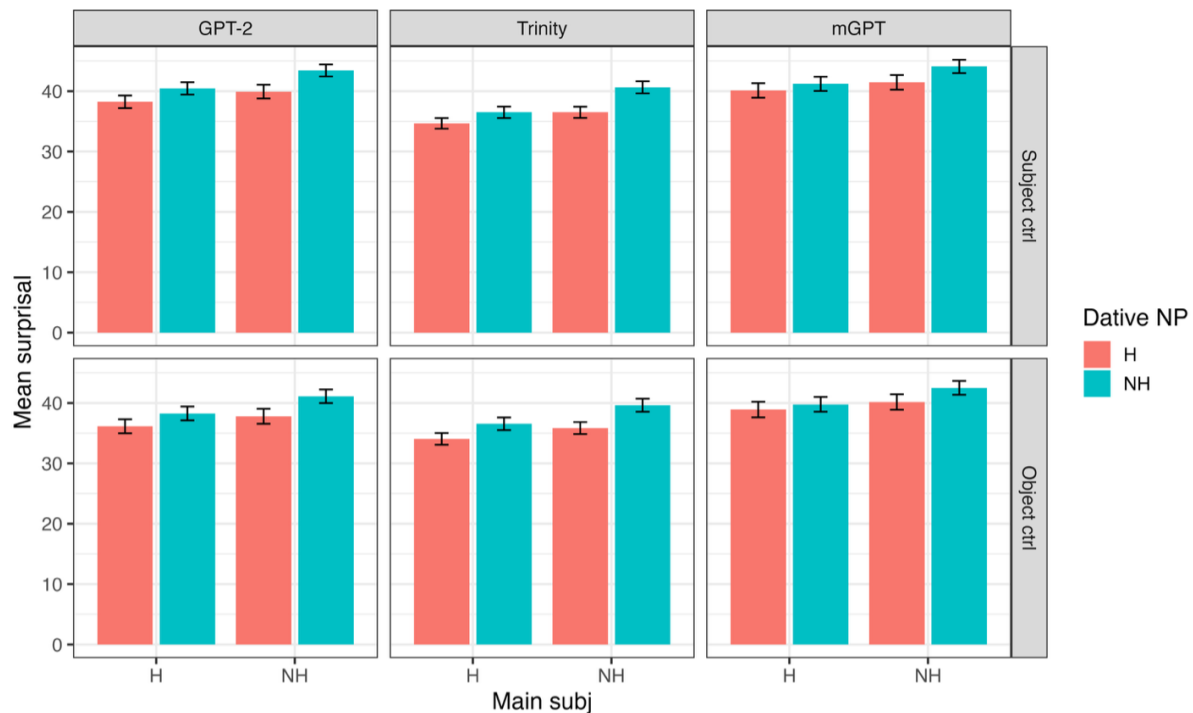


Figure 2: Mean surprisal values per condition within construction type by GPT language model. Error bars represent ± 1 standard error of the mean.

Table 3: Summary of likelihood ratio test results for GPT surprisal predicting human reading times in subject- and object-control constructions. * $p < .05$, ** $p < .01$, *** $p < .001$.

Construction	Measure	KoGPT-2	Ko-GPT-Trinity	mGPT
Subject control	First-pass	*	**	n.s.
	Regression-path	*	n.s.	n.s.
	Total Time	n.s.	**	n.s.
Object control	First-pass	n.s.	n.s.	*
	Regression-path	**	n.s.	**
	Total Time	*	*	***

was not significant after multiplicity adjustment ($p = .083$; Table B). More generally, the NH-H versus H-NH contrast was non-significant in every model and construction.

Critically, surprisal did not vary with distractor prominence in either construction. A combined analysis confirmed that the attraction effect did not interact with construction in any model (all $ps \geq .056$).

To evaluate whether the non-significant construction-dependent contrasts reflected genuine symmetry rather than insufficient power, we fitted Bayesian mixed-effects regressions in brms mirroring the combined frequentist analyses. First, Savage–Dickey Bayes factors favored the null

for the Construction $\times$ Main Subject $\times$ Dative NP interaction in all three models ($BF_{01} = 19.1$, 8.5, and 22.7 for KoGPT-2, Trinity, and mGPT), indicating moderate-to-strong evidence against construction-dependent modulation. Second, because our central prediction concerns the relative surprisal of the two mixed conditions (NH-H and H-NH), we tested whether the NH-H versus H-NH contrast differed across constructions. This contrast of contrasts was small and its credible interval spanned zero for every model (0.06 bits, 95% CrI $[-1.07, 1.19]$, $BF_{01} = 18.8$, KoGPT-2; -0.71 bits, $[-1.74, 0.32]$, $BF_{01} = 8.8$, Trinity; 0.14 bits, $[-0.86, 1.15]$, $BF_{01} = 21.4$, mGPT), and the within-construction NH-H versus

H-NH comparisons likewise favored the null ($BF_{01} = 4.5\text{--}19.2$ across models and constructions). By contrast, all three models showed evidence for the average Main Subject $\times$ Dative NP interaction itself ($BF_{10} = 5.4$ for KoGPT-2; > 100 for Trinity and mGPT), so the null-favoring results cannot be attributed to general insensitivity to the honorific manipulation.

In summary, all three models displayed a consistent surprisal hierarchy across the four conditions, yielding attraction-like effects. With the single exception of Trinity in object-control sentences, none of the models exhibited a clear grammaticality effect. More importantly, none showed evidence of prominence-driven modulation: the surprisal asymmetry between subject- and object-control constructions observed in human reading was absent in all three models.

5.3 Modeling human eye movements via computational surprisal

Table 3 summarizes statistical model outputs (see also Appendix D for complete analysis results). Language-model surprisal provided selective predictive value for human reading times, with efficacy varying by construction and eye-tracking measure. Note that improved fit reflects sensitivity to item-level difficulty rather than to prominence, which Section 5.2 showed the models lack.

For subject-control sentences, the predictive value of surprisal varied across models and processing measures. In first-pass reading time, both KoGPT-2 and Trinity significantly improved model fit over the baseline, while mGPT did not significantly improve fit ($p = .094$). For regression-path duration, only KoGPT-2 significantly improved fit, with Trinity and mGPT failing to contribute. For total reading time, Trinity significantly improved model fit, while KoGPT-2 and mGPT provided no predictive value. Thus, for subject-control structures, KoGPT-2 and Trinity showed measure-specific predictive value, with mGPT contributing little overall.

In object-control sentences, surprisal effects were observed across multiple measures. In first-pass reading time, only mGPT significantly improved baseline fit, with KoGPT-2 and Trinity failing to reach significance. In regression-path duration, KoGPT-2 and mGPT significantly improved model fit, whereas Trinity did not significantly improve fit ($p = .097$). In total reading time, all three models significantly improved fit. Overall, within

object-control sentences, mGPT showed the most consistent predictive value across measures, while KoGPT-2 contributed selectively and Trinity's effects were confined to total reading time.

Together, language-model surprisal selectively predicted human reading times. KoGPT-2 showed the broadest predictive value across constructions, mGPT was most reliable in object-control sentences, and Trinity contributed only selectively.

6 Discussion and Conclusion

In this study, we investigated whether GPT language models capture prominence-sensitive attraction in human sentence processing. We evaluated whether model surprisal exhibits prominence-driven modulation across Korean subject- and object-control constructions or operates symmetrically regardless of structural prominence. Human eye-tracking data and GPT surprisal diverged sharply. While human reading times showed a significant Construction $\times$ Attraction interaction across all measures ($ps < .045$), reflecting stronger interference from high-prominence subject distractors (Kwon and Sturt, 2019), none of the three GPT models reproduced this signature. Although model surprisal was sensitive to honorific mismatches, it did not interact with construction type. These dissociations suggest that human prominence sensitivity reflects the prominence-weighted retrieval constraints of memory-based parsing (Engelmann et al., 2019; Lewis and Vasishth, 2005) rather than expectation-based predictability (Hale, 2001; Levy, 2008). The results also suggest that LLMs align with human processing only where linear order coincides with structural prominence, and diverge where the two dissociate.

This divergence was systematic. Across models and constructions, surprisal scaled monotonically with the count of non-honorific elements, treating mixed conditions (H-NH and NH-H) symmetrically. The only deviation—a Trinity-specific contrast in object-control where H-NH incurred higher surprisal than NH-H—did not reflect structural encoding: it failed to generalize to subject-control, where the grammatical controller occupies the subject position. Its restriction to object-control, where the controller is also the NP nearest the critical verb, indicates a linear-proximity heuristic rather than sensitivity to prominence or grammatical role. Unlike human processing, where linear recency operates alongside prominence-based weighting

(Sturt and Kwon, 2024), the models showed no prominence-based weighting at all.

Decomposing dependency resolution into controller identification and prominence weighting clarifies where models fail. For controller identification, models exhibited a main effect of Construction (differentiating subject- and object-control clauses) but lacked significant interactions between Construction and the candidate NPs in the combined analysis ($ps \geq .13$), indicating that the detected clause-type difference did not produce construction-specific processing of the candidate NPs. For prominence weighting, the prominence account predicts stronger attraction from the high-prominence subject distractor in object-control (H-NH) than from the low-prominence object distractor in subject-control (NH-H); yet this difference did not emerge in any model ($ps \geq .056$; $BF_{01} \geq 8.8$). Thus, models neither used clause type to guide controller identification nor applied prominence-based weighting.

Crucially, higher predictive fit may be misleading: mGPT’s predictive advantage in object-control reading times stemmed from a general statistical penalty for feature incongruence rather than human-like prominence weighting. As English benchmarks often conflate linear and structural cues, apparent competence there may partly reflect coincidental alignments between linear order and structural roles (Chaves and Richter, 2021; Lasri et al., 2022). Control constructions break this alignment: the controller varies structurally while linear order is held constant, and Korean honorific agreement provides a socio-pragmatic dependency whose resolution requires identifying the structural controller. Our study found that prominence modulation was absent across model scales (125M–1.3B parameters) and training corpora, suggesting this pattern is not simply an artifact of model scale or one remedied by multilingual pretraining (Huang et al., 2024; Oh and Schuler, 2023). That is, while models track surface cues (Cuneo et al., 2025), they do not reflect the role-based prominence that governs human dependency resolution.

These findings collectively suggest that prominence-sensitive attraction does not fall out of next-token prediction over distributional data, and point to the need to incorporate processing mechanisms driven by working-memory constraints (e.g., memory activation and decay) into cognitive models of sentence processing, and potentially into architectures intended to capture human-like

dependency resolution. More broadly, they caution against treating computational surprisal as a direct proxy for human sentence-processing mechanisms, and underscore the value of evaluating models beyond English (Blasi et al., 2022; Chang and Bergen, 2024).

7 Limitations

Several limitations highlight directions for future research. First, our evaluation relied on the 320-instance item set from Kwon and Sturt (2019). While the human prominence effect is supported more broadly by meta-analysis (Jäger et al., 2017) and computational simulation (Engelmann et al., 2019), our model–human comparison is grounded in a single eye-tracking paradigm and one construction family; broader item sets, diverse clausal configurations, and additional human paradigms are needed to assess generalizability. Second, prominence was operationalized via grammatical roles in control structures. Because discourse prominence also encompasses topicality, animacy, and givenness, our findings identify a specific boundary condition: abstract, role-based asymmetries do not emerge organically in autoregressive architectures lacking human-like working-memory constraints (cf. Cuneo et al., 2025).

Methodologically, our analysis was restricted to decoder-only GPT variants using absolute positional encodings. Evaluating architectures with Rotary Position Embeddings (RoPE), bidirectional encoders, or instruction-tuned variants will clarify whether the absent prominence sensitivity stems from positional encoding schemes, autoregressive objectives, or broader model constraints.

Finally, our conclusions rest on output-level surprisal; structural-role information could be encoded internally yet remain unexpressed in surprisal, thereby limiting what the measure reveals rather than what the models represent. Targeted representation probing (e.g., diagnostic classification, representational similarity analysis) and head-level audits could determine whether such information is latently encoded despite its absence from surprisal.

References

- Mira Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.
- Geoff Bacon and Terry Regier. 2019. [Does bert agree? evaluating knowledge of structure depen-](#)

- dence through agreement relations. *arXiv preprint arXiv:1908.09892*.
- Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3).
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.
- Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.
- Wallace L. Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor, *Subject and topic: A new typology of language*, Proceedings of the Symposium on Subject and Topic (1975: University of California Santa Barbara), pages 25–55. Academic Press, New York, NY.
- Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.
- Rui P. Chaves and Stephanie N. Richter. 2021. Look at that! bert can be easily distracted from paying attention to morphosyntax. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 28–38. Association for Computational Linguistics.
- P. Contreras Kallens, R. D. Kristensen-McLachlan, and M. H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.
- N. Cuneo, N. Graves, S. Rakshit, and A. Goldberg. 2025. For gpt-4 as with humans: Information structure predicts acceptability of long-distance dependencies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.
- Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.
- Felix Engelmann, Lena A. Jäger, and Shravan Vasishth. 2019. The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, 43(12):e12800.
- Julie Franck, Saveria Colonna, and Luigi Rizzi. 2015. Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology*, 6.
- Michael C. Frank and Noah D. Goodman. 2026. Cognitive modeling using artificial intelligence. *Annual Review of Psychology*, 77:543–566.
- Talmy Givón. 1983. An introduction: Topic continuity in discourse. In *Topic Continuity in Discourse: A quantitative cross-language study*, pages 1–42. John Benjamins Publishing Company.
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001*, pages 1–8.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.
- K.-J. Huang, S. Arehalli, M. Kugemoto, C. Muxica, G. Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.
- Cameron R. Jones and Benjamin K. Bergen. 2024. People cannot distinguish gpt-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*.
- Lena A. Jäger, Felix Engelmann, and Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.
- Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82:1–26.
- Nayoung Kwon and Seongmin Mun. 2025. Evaluating syntactic generalization in transformer-based models using korean honorific agreement. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 26–37. Association for Computational Linguistics.

- Nayoung Kwon and Patrick Sturt. 2016. [Attraction effects in honorific agreement in Korean](#). *Frontiers in Psychology*, 7.
- Nayoung Kwon and Patrick Sturt. 2019. [Proximity and same case marking do not increase attraction effect in comprehension: Evidence from eye-tracking experiments in Korean](#). *Frontiers in Psychology*, 10.
- Sol Lago, Diego E. Shalom, Mariano Sigman, Ellen F. Lau, and Colin Phillips. 2015. [Agreement attraction in Spanish comprehension](#). *Journal of Memory and Language*, 82:133–149.
- Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. [Does BERT really agree? fine-grained analysis of lexical dependence on a syntactic task](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315. Association for Computational Linguistics.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Richard L. Lewis and Shravan Vasishth. 2005. [An activation-based model of sentence processing as skilled memory retrieval](#). *Cognitive Science*, 29(3):375–419.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). *arXiv preprint arXiv:1808.09031*.
- Brian McElree. 2000. [Sentence comprehension is mediated by content-addressable memory structures](#). *Journal of Psycholinguistic Research*, 29(2):111–123.
- Byung-Doh Oh and William Schuler. 2023. [Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?](#) *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Neal J. Pearlmutter, Susan M. Garnsey, and Kathryn Bock. 1999. [Agreement processes in sentence comprehension](#). *Journal of Memory and Language*, 41(3):427–456.
- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Soo Hyun Ryu and Richard Lewis. 2021. [Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Nathaniel J. Smith and Roger Levy. 2013. [The effect of word predictability on reading time is logarithmic](#). *Cognition*, 128(3):302–319.
- Patrick Sturt and Nayoung Kwon. 2024. [Agreement attraction in comprehension: Do active dependencies and distractor position play a role?](#) *Language, Cognition and Neuroscience*, 39(3):279–301.
- Julie A. Van Dyke and Brian McElree. 2006. [Retrieval interference in sentence comprehension](#). *Journal of Memory and Language*, 55(2):157–166.
- Shravan Vasishth, Bruno Nicenboim, Felix Engelmann, and Frank Burchert. 2019. [Computational models of retrieval processes in sentence processing](#). *Trends in Cognitive Sciences*, 23(11):968–982.
- Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. 2009. [Agreement attraction in comprehension: Representations and processes](#). *Journal of Memory and Language*, 61(2):206–237.
- Alex Warstadt and Samuel R. Bowman. 2020. [Can neural networks acquire a structural bias from raw linguistic data?](#) *arXiv preprint arXiv:2007.06761*.
- Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. [Testing the predictions of surprisal theory in 11 languages](#). *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- M. H. Yoo, J. Kim, and S. Song. 2025. [Multilingual capabilities of gpt: A study of structural ambiguity](#). *PLOS ONE*, 20(7):e0326943.

A Appendix

Table A: Fixed-effect estimates from linear mixed-effects models for GPT surprisal in subject-control constructions.

Model	Predictor	β	SE	t	p
KoGPT-2	(Intercept)	40.51	1.03	39.33	<.001
	Main Subject	−1.17	0.13	−8.85	<.001
	Dative NP	−1.43	0.13	−11.24	<.001
	Main Subject × Dative NP	0.32	0.07	4.29	<.001
Ko-GPT-Trinity	(Intercept)	37.07	0.92	40.18	<.001
	Main Subject	−1.49	0.12	−11.99	<.001
	Dative NP	−1.49	0.12	−12.30	<.001
	Main Subject × Dative NP	0.58	0.05	12.4	<.001
mGPT	(Intercept)	41.72	1.16	36.04	<.001
	Main Subject	−1.06	0.09	−11.40	<.001
	Dative NP	−0.94	0.13	−7.47	<.001
	Main Subject × Dative NP	0.38	0.09	4.12	<.001

Table B: Estimated marginal mean contrasts for the NH-H versus H-NH comparison, from the combined mixed-effects models. Estimates are differences in surprisal (bits; NH-H minus H-NH); 95% confidence intervals are Bonferroni-adjusted and *p*-values are Holm-adjusted across the six pairwise comparisons within each model. The contrast was non-significant in every model × construction combination.

Model	Construction	Estimate	SE	95% CI	<i>p</i>
KoGPT-2	Subject control	−0.53	0.46	[−1.77, 0.71]	.252
	Object control	−0.48	0.46	[−1.71, 0.76]	.302
Ko-GPT-Trinity	Subject control	0.00	0.41	[−1.09, 1.10]	.999
	Object control	−0.71	0.41	[−1.81, 0.38]	.083
mGPT	Subject control	0.25	0.40	[−0.82, 1.31]	.540
	Object control	0.38	0.40	[−0.68, 1.45]	.337

Table C: Fixed-effect estimates from linear mixed-effects models for GPT surprisal in object-control constructions.

Model	Predictor	β	SE	t	p
KoGPT-2	(Intercept)	38.33	1.15	33.42	<.001
	Main Subject	−1.12	0.13	−8.65	<.001
	Dative NP	−1.36	0.11	−12.64	<.001
	Main Subject × Dative NP	0.30	0.11	2.80	.006
Ko-GPT-Trinity	(Intercept)	36.52	1	36.35	<.001
	Main Subject	−1.22	0.14	−9.00	<.001
	Dative NP	−1.57	0.13	−12.36	<.001
	Main Subject × Dative NP	0.32	0.04	7.28	<.001
mGPT	(Intercept)	40.34	1.22	32.95	<.001
	Main Subject	−0.99	0.09	−11.33	<.001
	Dative NP	−0.80	0.13	−5.95	<.001
	Main Subject × Dative NP	0.37	0.09	4.25	<.001

Table D: Likelihood ratio test results comparing baseline and surprisal-augmented models for subject- and object-control constructions.

	Measure	Model comparison	χ^2	df	<i>p</i>
Subject control	First-pass	Baseline + KoGPT-2	4.60	1	.032*
		Baseline + Trinity	8.12	1	.004**
		Baseline + mGPT	2.80	1	.094
	Regression-path	Baseline + KoGPT-2	4.09	1	.043*
		Baseline + Trinity	2.49	1	.115
		Baseline + mGPT	1.44	1	.23
	Total Time	Baseline + KoGPT-2	3.74	1	.053
		Baseline + Trinity	7.23	1	.007**
		Baseline + mGPT	2.23	1	.135
Object control	First-pass	Baseline + KoGPT-2	2.46	1	.117
		Baseline + Trinity	2.73	1	.098
		Baseline + mGPT	5.53	1	.019*
	Regression-path	Baseline + KoGPT-2	6.99	1	.008**
		Baseline + Trinity	2.76	1	.097
		Baseline + mGPT	7.80	1	.005**
	Total Time	Baseline + KoGPT-2	6.47	1	.011*
		Baseline + Trinity	5.30	1	.021*
		Baseline + mGPT	12.17	1	<.001***

Table E: Random-effects structure for every regression. Main Subject and Dative NP are sum-coded honorific factors; “Intercept” denotes a random intercept. Surprisal models include only by-item terms (surprisal values have no by-participant variation); reading-time models include by-participant and by-item terms. For each reading-time measure, the baseline and all surprisal-augmented models share the listed structure.

Analysis	Model / Measure	Random-effects structure
<i>Surprisal, subject control</i>		
	KoGPT-2	(1 + Main Subject + Dative NP item)
	Ko-GPT-Trinity	(1 + Main Subject + Dative NP item)
	mGPT	(1 + Dative NP item)
<i>Surprisal, object control</i>		
	KoGPT-2	(1 + Main Subject item) ^a
	Ko-GPT-Trinity	(1 + Main Subject + Dative NP item)
	mGPT	(1 + Dative NP item)
<i>Surprisal, combined (Construction × Main Subject × Dative NP)</i>		
	KoGPT-2	(1 + Main Subject + Dative NP item)
	Ko-GPT-Trinity	(1 + Main Subject + Dative NP item)
	mGPT	(1 + Dative NP item)
<i>Reading times, subject control</i>		
	First-pass	(1 + Dative NP subj) + (1 item)
	Regression-path	(1 subj) + (1 + Main Subject item)
	Total time	(1 subj) + (1 + Main Subject item)
<i>Reading times, object control</i>		
	First-pass	(1 subj) + (1 + Main Subject item)
	Regression-path	(1 subj) + (1 item)
	Total time	(1 + Dative NP subj) + (1 + Dative NP item)

^aThe full structure (1 + Main Subject + Dative NP | item) was singular; the model was refit with (1 + Main Subject | item), yielding virtually identical estimates.