What GPT language models get right—and wrong—about Korean subject honorification

Gyu-Ho Shin (ghshin@uic.edu)
Department of Linguistics, University of Illinois Chicago
601 South Morgan Street, Chicago, IL 60607, USA

Nayoung Kwon (nkwon@uoregon.edu)
Department of East Asian Languages, University of Oregon
1030 East 13th Ave, Eugene, OR 97403-1245, USA

Seongmin Mun (seongminmun@knu.ac.kr)
Department of English Language and Literature, Kyungpook National University
80, Daehak-ro, Buk-gu, Daegu 41566, South Korea

Abstract

The present study investigates how GPT language models approximate human sentence-processing dynamics observed in Korean subject honorification. Drawing upon two open-access datasets, we compare human responses with surprisal estimates from two Korean-capable GPT models. Across tasks, the models reliably identified overt morphosyntactic violations, assigning high surprisal to clear honorific agreement mismatches associated with low politeness ratings and increased reading times in humans. However, GPT surprisal failed to capture more graded and context-sensitive effects, including the optionality of the subject honorific suffix, animacy-dependent politeness evaluations, and delayed spillover effects in human reading. Correlations between human measures and model surprisal were consistently weak and highly condition-specific. Taken together, while GPT surprisal serves as a coarse indicator of well-formedness, it inadequately represents the socio-pragmatic and discourse-level constraints that shape Korean sentence processing. This calls for the need for more cognitively and culturally grounded metrics in language sciences and explainable AI.

Keywords: GPT; surprisal; sentence processing; pragmatic competence; Korean subject honorification; explainable AI

Introduction

A prominent trend in the language sciences is the application of computational methods to linguistic inquiry. This line of research has examined the extent to which computational models simulate human language behaviour (Frank & Goodman, 2025; Hawkins et al., 2020; Jones & Bergen, 2024; Marvin & Linzen, 2019; Warstadt et al., 2019; Wilcox et al., 2018), as well as performance differences across algorithms (Hu et al., 2020; Shin & Mun, 2025). These approaches have gained traction in efforts to explain the cognitive mechanisms underlying human language processing (Contreras Kallens et al., 2023; Warstadt & Bowman, 2020). With ongoing technological advances, computational models increasingly complement traditional paradigms by revealing general information-processing principles in sentence comprehension. Recent work further assesses whether insights from computational simulations converge with behavioural and corpus-based evidence concerning core properties of human language (Ambridge et al., 2020; Oh et al., 2022; Shin & Mun, 2025; Xu et al., 2023). Despite this progress, the field remains heavily skewed towards a small set of languages, particularly English (Blasi et al., 2022; Chang & Bergen, 2024). This imbalance is reinforced by the predominance of English-centred Large Language Models (LLMs), limiting the generalisability of existing findings to typologically distinct and under-resourced languages. Consequently, progress towards explainable AI (XAI)—interpretable evaluations of the extent to which model behaviour plausibly reflects human language processing—remains constrained.

The present study addresses these limitations by examining how effectively Transformer-based language models capture human sentence-processing dynamics in Korean subject honorification, a system governed by both structural and socio-pragmatic conventions. Transformers are deep neural architectures that construct contextual representations through self-attention, enabling the parallel modelling of long-range dependencies. When pretrained at scale, as in the Generative Pre-trained Transformer (GPT; Radford et al., 2018), they generate probabilistic predictions over upcoming words. We adopt GPT as a representative decoder-only Transformer in which token embeddings and positional encodings are processed through stacked multi-head attention and feed-forward layers, with residual connections and layer normalisation enabling conditioning on all prior context (Radford et al., 2018; Vaswani et al., 2017). Trained via next-token prediction on large corpora, GPT functions as a general-purpose learner without task-specific supervision and supports zero- and few-shot generalisation through prompting (Brown et al., 2020; Radford et al., 2019). Model-derived probabilities and surprisal have been shown to correlate with acceptability judgements and processing measures (e.g. Schrimpf et al., 2021; Warstadt et al., 2020), and recent work reports further links between GPT's grammatical knowledge and behavioural data (e.g. Goldstein et al., 2022; Yoo et al., 2025). Notably, GPT-4 demonstrates sensitivity to information structure, such that prominence

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

manipulations predict and causally influence acceptability in long-distance dependencies (Cuneo et al., 2025).

Across GPT-based evaluations, two recurring patterns emerge. In grammaticality and acceptability judgements, GPT-2 reliably tracks human preferences in clausal alternations, capturing verb biases in argument structure with high accuracy (Hawkins et al., 2020). By contrast, GPT-3 variants and ChatGPT often yield mixed and affirmation-biased judgements that are highly sensitive to prompting, reducing their reliability as grammaticality evaluators (Dentella et al., 2023; Lee & Wang, 2023). With respect to reading times, surprisal robustly predicts processing difficulty, typically following a logarithmic predictability–cost relationship. Smaller GPT-2 models often provide superior fits, whereas larger models may underperform and spuriously favour super-logarithmic patterns (de Varda et al., 2024; Shain et al., 2024). Nevertheless, GPT-2 surprisal underestimates garden-path effects and fails to capture main-verb slowdowns in deeply embedded structures, indicating persistent divergence from human processing (Oh & Schuler, 2023). Large-scale benchmarks further show that model surprisal underestimates syntactic ambiguity, mischaracterises relative difficulty and explains limited item-level variance, suggesting that next-word prediction alone is insufficient to model human sentence processing (Huang et al., 2024). Alternative sentence- and discourse-level measures may therefore better capture aspects of comprehension and cross-linguistic variation (Sun & Wang, 2025; Zhou et al., 2025). Overall, GPT-family models exhibit promising yet qualified correspondence with human judgements and processing dynamics, with alignment strongly modulated by model size and training regime.

Korean, which is typologically distant from English, remains substantially underrepresented in this literature, limiting the generalisability of English-based findings. To address this gap, the present study focuses on subject honorification, a subject–predicate dependency that differs markedly from English in both structure and function. While native speakers show strong sensitivity to violations of honorific feature agreement, actual usage diverges from canonical agreement systems such as number or person in English. By comparing model-derived probabilistic expectations with established behavioural evidence, we evaluate whether GPT predictions align with human sentence-processing dynamics in a typologically distinct context, contributing to a more explainable account of model–human correspondence.

Subject honorification in Korean

Korean is an agglutinative Subject–Object–Verb language that encodes grammatical relations through extensive case marking and verbal morphology. It is highly context-dependent and situation-oriented, allowing linguistic elements to be omitted when pragmatically recoverable. The honorification system, which encodes the speaker's stance towards a referent, requires fine-grained form–context mappings to achieve communicative goals. Crucially, it integrates grammatical and socio-cultural knowledge (Brown, 2015; Sohn, 1999), rendering it central to communicative competence in Korean.

Korean subject honorification consists of two core components: an honorifiable subject and the subject honorific suffix *-(u)si* (SUBJ.HON) attached to the predicate stem. In addition, speakers encode deference towards the addressee through six addressee-honorification levels, realised via sentence-final verbal endings (Sohn, 1999). In (1), honorific agreement is established between the subject (*sensayngnim* 'teacher') and the predicate marked with SUBJ.HON, while politeness towards the addressee is expressed through the sentence ender *-eyo/ayo*.

(1) Korean subject honorification
Sensayngnim, sensayngnim-kkeyse
Teacher, teacher-NOM.HON
wusung-ul ha-si-ess-eyo.
championship-ACC do-SH-PST-DCL.POL[1]
'Teacher, you won the championship.'

Previous research on subject honorification processing demonstrates that native Korean speakers are sensitive to honorific feature agreement. The use of SUBJ.HON with a non-honorifiable subject (e.g. *kkoma* 'child'), as in (2), induces processing difficulty, evidenced by increased reading times (Kwon & Sturt, 2016, 2019) and P600 effects (Kwon & Sturt, 2024)—a neural correlate commonly associated with number and person agreement violations across languages (e.g. Hinojosa et al., 2003; Osterhout & Mobley, 1995).

(2) Non-honorifiable subject + SUBJ.HON
*kkoma-ka wusung-ul ha-si-ess-eyo.
*kid-NOM championship-ACC do-SH-PST-DCL.POL
'The kid won the championship.'

Unlike English number or person agreement, subject honorification in Korean is highly context-sensitive. The omission of SUBJ.HON in the presence of an honorifiable subject does not necessarily render a sentence ungrammatical, as in (3).

(3) Honorifiable subject + no SUBJ.HON
sensayngnim-kkeyse wusung-ul hay-ess-eyo.
teacher-NOM.HON championship-ACC do-PST-DCL.POL
'The teacher won the championship.'

Such flexibility arises when formality or hierarchical distance between interlocutors is reduced, or when the speaker–addressee relationship is sufficiently close (Choo & Kwak, 2008; Sohn, 1999). The high optionality of SUBJ.HON (Kim et

[1] Abbreviations: ACC = accusative case marker; AH = addressee honorific marker; COMP = complementiser; DCL = ending, declarative; DFL = speech level, deferential; DIR = directional marker; H = honorific; NH = non-honorific; NOM = nominative case marker; POL = speech level, polite; PST = past tense marker; SH = subject honorific marker; TOP = topic marker.

al., 2005; Song et al., 2019) therefore indicates that an honorifiable subject does not consistently require its use.

Moreover, subject honorification may extend to contextually associated inanimate entities (e.g. *khephi* 'coffee'), as in (4), despite their intrinsically non-honorifiable status—a phenomenon known as *inanimate honorification* (Brown, 2022; Choo & Kwak, 2008).

(4) Inanimate honorification
Sonnim, khephi nao-si-ess-supnita
Customer.HON, coffee come.out-SH-PST-DCL.DFR
'Customer, your coffee is ready.'

This usage is increasingly attested in naturalistic settings, particularly in customer service interactions, where it functions to express respect towards addressees with situational authority (Eom, 2019). While some researchers classify inanimate honorification as ungrammatical (Kim-Renaud, 2001), others interpret it as a politeness strategy reflecting speakers' sensitivity to social context (Kwon & Lee, 2024). Such sensitivity also appears to shape native speakers' online processing of SUBJ.HON (Kim & Shin, 2021).

Despite the suitability of Korean subject honorification as a test case for evaluating human-like linguistic competence in LLMs, research in this domain remains sparse. Two closely related studies provide indirect evidence. Kwon and Mun (2026), focusing directly on subject honorification, found that GPT-2 outperformed BERT in classification accuracy, although both architectures remained sensitive to surface features rather than demonstrating genuine syntactic competence. Lee and Wang (2023), examining addressee honorification, evaluated GPT-2 and BERT-family models on politeness dependencies between politeness-sensitive pronouns and polite sentence-final endings, finding performance at or below chance. Critically, neither study directly compared model behaviour with online human sentence-processing data, leaving open the question of whether LLMs approximate the incremental dynamics observed in native speakers.

This study therefore addresses the following research question: to what extent do GPT language models simulate adult native Korean speakers' sentence-processing patterns involving subject honorification? To this end, we draw upon open-access datasets from two tasks—politeness rating and self-paced reading—reported in Kwon and Shin (2025), and evaluate three GPT variants. Each model is assessed along two dimensions. First, we measure accuracy in judgements of grammaticality or appropriateness using human rating data. Second, we estimate region-by-region reading times to assess alignment with human processing dynamics, drawing on self-paced reading and eye-tracking benchmarks. The results are expected to delineate both the strengths and limitations of LLMs in simulating human sentence processing, while clarifying the roles of language-specific properties, task demands and architectural constraints.

Methods

Dataset

Politeness Rating (native-speaker subset; $n = 20$). Politeness ratings were collected for 216 sentences (36 per condition) by crossing (i) subject type—honorifiable, non-honorifiable, and inanimate (e.g. *kimchi*)—and (ii) the presence versus absence of SUBJ.HON on the predicate. Non-honorifiable subjects were included to induce honorific mismatch, and inanimate subjects to induce semantic incongruity. Sentences were rated on a six-point Likert scale (0 = *very impolite*, 5 = *very polite*). Responses with reaction times below 1,000 ms or above 10,000 ms were excluded (approximately 10% data loss), and the trimmed dataset was used for analysis.

Self-paced Reading (native-speaker subset; $n = 40$). This dataset employed the same sentences under identical experimental conditions, differing only in word order. In the politeness-rating materials, subjects preceded predicates within a simple clause (e.g. *teacher*-NOM *classroom*-DIR *enter*-SH-PST-DECL). In the self-paced reading materials, predicates appeared in Region 3 and subjects in Region 4 across clausal boundaries (e.g. *classroom*-DIR *enter*-SH-while *teacher*-NOM …). Example stimuli are provided in (5).

(5) Example sentences for the self-paced reading task in Kwon and Shin (2025): 'Hastily entering the classroom, the teacher/Mia/the picture made a loud noise.'

(a) subject (honorifiable, non-honorifiable, inanimate) + predicate with SUBJ.HON
[kyosil-lo]$_{R1}$ [kuphi]$_{R2}$ [tuleka-**si**-mye]$_{R3}$
classroom-dir hastily enter-**sh**-while
[sensayngnim-kkeyse/Mia-ka/kulim-i]$_{R4}$
teacher-nom.hon/Mia-nom/picture-nom
[khun]$_{R5}$ [soli-lul]$_{R6}$ [nay-ss-eyo]$_{R7}$
big sound-acc make-pst-dcl

(b) subject (honorifiable, non-honorifiable, inanimate) + predicate without SUBJ.HON
[kyosil-lo]$_{R1}$ [kuphi]$_{R2}$ [tuleka-mye]$_{R3}$
classroom-dir hastily enter-while
[sensayngnim-kkeyse/Mia-ka/kulim-i]$_{R4}$
teacher-nom.hon/Mia-nom/picture-nom
[khun]$_{R5}$ [soli-lul]$_{R6}$ [nay-ss-eyo]$_{R7}$
big sound-acc make-pst-dcl

Raw reading times were trimmed by excluding values below 150 ms or above 4,000 ms and observations exceeding three standard deviations from the condition mean (approximately 2% data loss). The remaining data were log-transformed and residualised to control for word length and individual reading speed (Baayen & Milin, 2010). Following Trueswell et al. (1994), predicted reading times were estimated for each participant using a linear mixed-effects model with region word length (syllable count) as a fixed effect and participant as a random effect. Residual reading times were computed as the difference between observed and predicted values. The primary region of interest was Region

4—the subject region immediately following the (non-)honorific predicate—where honorific agreement between subject and predicate, including animacy, is computed.

Modelling

We employed two open-access GPT variants available for Korean (Table 1): KoGPT-2[2] (GPT-2 compatible) and Ko-GPT-Trinity [3] (GPT-3 compatible). All modelling was conducted on a MacBook Pro (Apple M4 Max, 16-core CPU, 40-core GPU, 16-core Neural Engine, 128GB memory).

Our workflow involved loading two models from Hugging Face, standardising tokeniser behaviour, and processing sentences in plain-text format (one sentence per line). Each sentence was segmented into *ejels*—a Korean-specific spacing unit approximating a word—using whitespace, after which each ejel was further tokenised into subword units by the model-specific tokeniser. Subword tokens were mapped to numerical identifiers and assembled into input sequences, optionally preceded by a beginning-of-sentence token, before being passed to the model to obtain next-token probability distributions. For each model, probabilities were computed conditional on preceding context, and surprisal was calculated as the negative logarithm of these probabilities (Levy, 2008). Ejel-level surprisal was obtained by summing over constituent subwords, and sentence-level surprisal by aggregating across ejels, yielding fine-grained estimates of processing difficulty.

Table 1. Model specifications

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

We used surprisal in relation to politeness ratings not as a direct measure of politeness itself, but as an index of how expected each form is under the model's learnt distribution. If a model has internalised the socio-pragmatic constraints underlying honorific use, items that violate those constraints should elicit higher surprisal, corresponding to lower politeness ratings. The relationship between the two therefore provides an empirical test of whether the models' probability distributions align with human socio-pragmatic intuitions, extending surprisal-based evaluation to socio-pragmatic understanding—beyond its more typical applications in modelling incremental processing and acceptability.

Statistical Analysis

The politeness-rating data were analysed using cumulative link mixed models with ordinal regression, implemented in the ordinal package (Christensen, 2023) in R (R Core Team, 2025). A logit link function modelled cumulative probabilities of higher ratings. Model diagnostics evaluated convergence, the proportional-odds assumption and residual distributions. Effect sizes were approximated using Nagelkerke's R^2, computed by comparing the log-likelihood of the final model with that of a null model containing only a fixed intercept and the full random-effects structure. Pairwise post hoc comparisons were conducted using the emmeans package (Lenth, 2025), with p-values obtained from Wald *z*-tests as implemented in ordinal (Christensen, 2023).

The self-paced-reading data were analysed using linear mixed-effects models fitted separately for each critical and spillover region with the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2025). Model fit was summarised using Nakagawa's conditional R^2 (Nakagawa & Schielzeth, 2013). GPT-derived surprisal values were analysed with separate linear mixed-effects models for each GPT variant, using the same fixed-effects structure but including Item as the sole random effect. All other analytical settings matched those used for the human rating and reading-time analyses. Correspondence between human and GPT data was assessed by computing Pearson's *r* for each GPT variant within each dataset.

Results and Discussion

Politeness Rating

If GPT surprisal aligns with human politeness judgements, two main patterns are expected. First, surprisal should be lowest when politeness cues are fully realised—namely, when an honorifiable subject co-occurs with an honorific predicate suffix. Second, surprisal should increase gradationally with cue misalignment, particularly with respect to subject honorifiability, as established in the original study (Kwon & Shin, 2025). Sentences with inanimate subjects are therefore predicted to elicit higher surprisal.

Figure 1 presents human politeness ratings alongside GPT-derived surprisal values. Human participants rated the h_h condition as most polite, conditions with non-honorifiable animate/human subjects as less polite, and conditions with inanimate subjects as least polite. GPT models, however, exhibited a different pattern: among animate/human subjects, the nh_h condition elicited the highest surprisal, while surprisal for the inanimate nh_nh condition was comparable to that of its animate counterpart. These patterns were confirmed statistically. Human ratings showed main effects of *Subject* ($p < 0.001$) and *Predicate* ($p = 0.001$), as well as a significant *Subject* × *Predicate* interaction ($p < 0.001$). Bonferroni-corrected post-hoc tests revealed reliable differences between each inanimate condition and its animate/human counterpart (all ps < 0.001).

The GPT models did not reproduce this pattern. Although all three models showed main effects of Subject and Predicate and a significant interaction (all ps < 0.001), post-hoc comparisons revealed no reliable differences between

[2] https://huggingface.co/skt/kogpt2-base-v2

[3] https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5

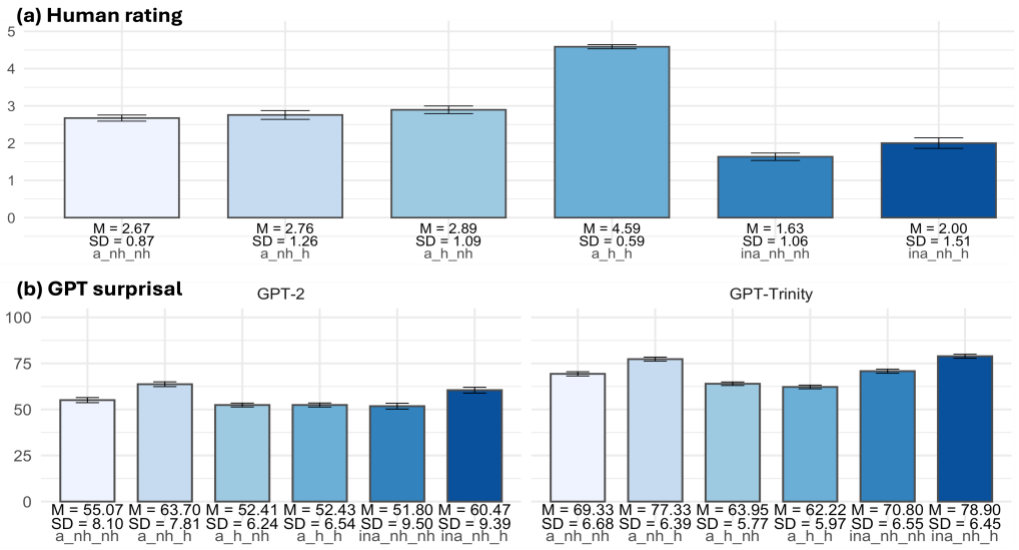

Figure 1. Politeness rating. X-axis: Conditions; Y-axis: (a) Rating, (b) Surprisal. Error bars indicate 95% CIs (SE).

corresponding animate/human and inanimate conditions (a_nh_nh vs. ina_nh_nh: $p = 0.139$ for Ko-GPT2, $p = 0.775$ for Ko-GPT-Trinity; a_nh_h vs. ina_nh_h: $p = 0.106$ for Ko-GPT2, $p = 1$ for Ko-GPT-Trinity). Correlations between human politeness ratings and GPT surprisal were weak and did not reach statistical significance.

In sum, human politeness judgements were strongly modulated by distinctions between animate/human and inanimate subjects, whereas GPT models failed to reflect this sensitivity. Instead, they assigned highest surprisal to the animate/human nh_h condition and treated inanimate conditions similarly to their non-honorifiable animate counterparts. Together with the absence of reliable correlations, these findings indicate limited alignment between GPT surprisal and human politeness judgements in subject honorification.

Self-paced reading

If GPT surprisal aligns with human self-paced reading, three patterns are expected. First, surprisal at R4 (the subject region) should be higher in feature-mismatching conditions (animate/human nh_h, inanimate nh_h) than in feature-matching conditions (animate/human nh_nh, inanimate nh_nh). Second, despite the absence of SUBJ.HON at R3 (the predicate region), the animate/human h_nh condition should elicit lower surprisal at R4 than other mismatching conditions, reflecting the optionality of SUBJ.HON. Third, for non-honorifiable subjects (animate/human nh_h, inanimate nh_h), animacy should not substantially modulate surprisal, yielding comparable values across these conditions.

Figure 2 presents region-by-region human reading times alongside GPT-derived surprisal. In the human data, R4—where honorific agreement with the preceding predicate is computed—showed the greatest processing cost in the inanimate nh_h condition, followed by the animate/human nh_h condition, indicating increased difficulty arising from feature mismatches between honorific predicates and non-honorifiable subjects. GPT surprisal, however, followed a different pattern: across models, surprisal at R4 was lowest in the animate/human h_h condition and highest in the animate/human nh_nh and nh_h conditions, both of which involve non-honorifiable subjects irrespective of predicate marking.

These patterns were supported by statistical analyses. For human reading times at R4, there were main effects of *Subject* ($p = 0.013$) and *Predicate* ($p = 0.011$), as well as a significant *Subject* × *Predicate* interaction ($p = 0.018$). Bonferroni-corrected post-hoc tests showed longer reading times (i) in the inanimate nh_h condition than in the animate/human h_h ($p = 0.008$) and h_nh conditions ($p = 0.008$), and (ii) in honorific- than non-honorific-predicate conditions within inanimate subject items ($p = 0.012$). These findings

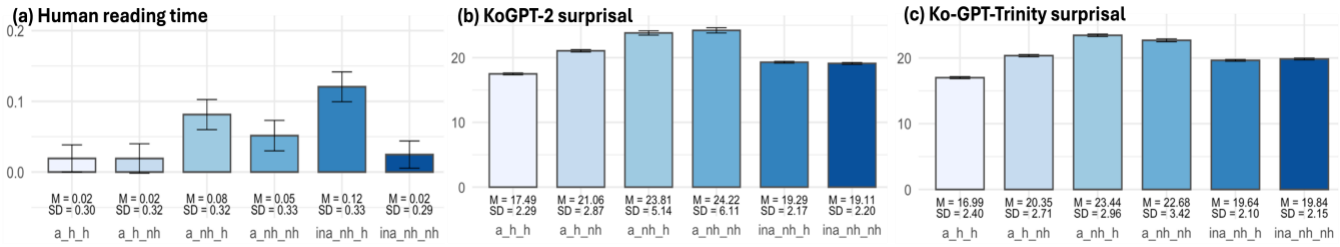

Figure 2. Self-paced reading (R4). X-axis: Condition; Y-axis: (a) Residual reading time, (b–d) Surprisal. Error bars indicate 95% CIs (SE).

demonstrate sensitivity to subject animacy and to the optional, context-dependent use of honorific predicates.

By contrast, GPT surprisal at R4 revealed a main effect of *Predicate* moderated by a *Subject* × *Predicate* interaction across all models (all ps < 0.001), but post-hoc tests showed no reliable effects in conditions where human readers exhibited robust differences (all ps > 0.1). Instead, models consistently produced higher surprisal (i) for animate/human nh_h than for inanimate nh_h conditions (all ps < 0.05), (ii) for animate/human nh_nh than for inanimate nh_nh conditions (all ps < 0.01), and (iii) for non-honorific- than honorific-predicate conditions within animate honorifiable subjects (a_h_nh vs. a_h_h; all ps < 0.001). These results indicate limited model sensitivity to honorific agreement involving subject animacy at this region.

Finally, alignment between human reading times and GPT surprisal was assessed. As in the politeness-rating analyses, none of the correlations reached statistical significance. Overall, the reading-time results reveal weak and unreliable alignment between human processing and GPT-derived surprisal, mirroring the patterns observed for politeness judgements.

Discussion and Conclusion

Human judgements and model-derived surprisal diverged markedly. Human participants exhibited a clear pragmatic hierarchy: sentences combining an honorific suffix with an honorifiable animate/human subject (a_h_h) were rated most polite; sentences with animate subjects but honorific discord were judged less polite; and sentences with inanimate subjects were rated least polite, reflecting the intuition that honorifying inanimate entities is inappropriate in neutral contexts. Human ratings were therefore sensitive to both morphosyntactic marking and subject animacy. The GPT models failed to reproduce this gradient. Instead, surprisal was driven primarily by overt syntactic anomaly: across both animate/human and inanimate categories, non-honorifiable subjects paired with an honorific suffix (nh_h) elicited the highest surprisal, consistent with detection of a grammatical violation. Crucially, the models did not encode animacy-based politeness distinctions, as inanimate items judged highly impolite by humans yielded surprisal values comparable to their animate counterparts. Statistical analyses confirmed that none of the models produced reliable surprisal differences between corresponding animate/human and inanimate conditions. We interpret this as evidence of limited socio-pragmatic integration in next-token prediction models: while they respond to formal honorific violations, they do not represent whether a referent is socially appropriate for honorification. This extends prior findings on LLMs' pragmatic limitations (e.g., Cong, 2024; Ruis et al., 2023; Sravanthi et al., 2024) to Korean politeness phenomena.

Human readers likewise showed a pronounced slowdown at Region 4 (the critical subject region) when an honorific predicate preceded a non-honorifiable subject, with the strongest effect for inanimate mismatches (ina_nh_h) and a smaller but robust slowdown for animate mismatches (a_nh_h). These results align with evidence that inappropriate honorific use in Korean induces processing difficulty akin to agreement violations. GPT models also showed elevated surprisal for non-honorifiable subjects, assigning lowest surprisal to fully feature-matching sentences and higher surprisal whenever the subject was non-honorifiable. At this coarse level, the models appeared to mirror human sensitivity to mismatch. However, closer inspection revealed systematic misalignment. Human slowdowns were specifically triggered by the combination of an honorific cue and a non-honorifiable subject, whereas the models often treated non-honorifiable subjects as intrinsically surprising, even in grammatical contexts. This pattern suggests an over-generalised preference for honorifiable subjects rather than context-sensitive agreement computation. Animacy effects further diverged: all models showed larger surprisal increases for animate mismatches than for inanimate ones, opposite to the dominant human trend. Overall, GPT surprisal captured some aspects of processing difficulty but frequently attributed difficulty to different underlying factors, limiting its interpretability as a proxy for human parsing mechanisms.

Taken together, the Korean subject-honorification results both converge with and depart from patterns reported in earlier, predominantly English-based evaluations. Convergently, GPT surprisal tracks overt grammatical anomalies in a recognisably human-like manner: clear honorific agreement violations are penalised by both models and human readers, consistent with predictability–cost accounts. The departures, however, are systematic and theoretically consequential. As prior English studies have noted, next-word prediction alone does not capture the full complexity of human sentence processing; in Korean, these limitations are amplified. The models failed to capture graded, context-sensitive acceptability and largely ignored how animacy and social appropriateness shape politeness judgements—precisely the interface where grammar and pragmatics interact. Surprisal thus often signals that processing is difficult without capturing why it is difficult for human comprehenders.

From a cross-linguistic perspective, these findings underscore a broader language bias in the field. Heavy reliance on English may overstate model–human alignment, whereas typologically and culturally distinct languages such as Korean expose weaknesses in contextual and pragmatic modelling. For explainable AI, this implies that superficial correspondence between surprisal and behavioural data may conceal mechanistic divergence, motivating richer modelling objectives and evaluation frameworks beyond token-level prediction. Honorification, which encodes social stance in addition to morphosyntax, provides a particularly revealing test case: while the models captured the formal rule, they failed to acquire its semantic–pragmatic calibration. Overall, GPT-derived surprisal emerges as a useful but incomplete proxy for processing cost—effective for detecting clear anomalies, yet explanatorily fragile when comprehension depends on pragmatic appropriateness, cultural norms, or discourse-level inference.

References

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., Bannard, C., Samanta, S., McCauley, S., Arnon, I., Bekman, D., Efrati, A., Berman, R., Narasimhan, B., Sharma, D. M., Nair, R. B., Fukumura, K., Campbell, S., Pye, C., Pixabaj, S. F. C., Paliz, M. M., & Mendoza, M. J. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. *Cognition, 202*, 104310. https://doi.org/10.1016/j.cognition.2020.104310

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Brown, L. (2015). Honorifics and politeness. In L. Brown & J. Yeon (Eds.), *The handbook of Korean linguistics* (pp. 303–319). Oxford: John Wiley & Sons, Inc.

Brown, L. (2022). Politeness as normative, evaluative and discriminatory: the case of verbal hygiene discourses on correct honorifics use in South Korea. *Journal of Politeness Research*, *18*(1), 63–91. https://doi.org/10.1515/pr-2019-0008

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. https://arxiv.org/abs/2005.14165

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, 1–58. https://doi.org/10.1162/coli_a_00492

Choo, M. & Kwak, H-Y. (2008). Using Korean: a guide to contemporary usage. Cambridge: Cambridge University Press.

Cong, Y. (2024). Manner implicatures in large language models. *Scientific Reports*, *14*(1), 29113. https://doi.org/10.1038/s41598-024-80571-3

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Christensen, R. H. B. (2023). *Ordinal-regression models for ordinal data* (R package version 2023.12-4). https://CRAN.R-project.org/package=ordinal

Cuneo, N., Graves, N., Rakshit, S., & Goldberg, A. (2025). For GPT-4 as with Humans: Information Structure Predicts Acceptability of Long-Distance Dependencies. *Proceedings of the Annual Meeting of the Cognitive Science Society, 47*. Retrieved from https://escholarship.org/uc/item/6rk7s243

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51), e2309583120. https://doi.org/10.1073/pnas.2309583120

Eom, J-s. (2019). A study on the use of the honorific suffix '-*si*-' by learners of Korean. *Hanmincokemwunhak*, 84, 9–50.

Frank, M. C., & Goodman, N. D. (2025). Cognitive modeling using artificial intelligence. *Annual Review of Psychology*, *77*. https://doi.org/10.1146/annurev-psych-030625-040748

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hawkins, R. D., Yamakoshi, T., Griffiths, T. L., & Goldberg, A. E. (2020). Investigating representations of verb bias in neural language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4653–4663). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2010.02375

Hinojosa, J., Martín-Loeches, M., Casado, P., Munoz, F., & Rubia, F. (2003). Similarities and differences between phrase structure and morphosyntactic violations in Spanish: An event-related potentials study. *Language and Cognitive Processes*, *18*(2), 113–142. https://doi.org/10.1080/01690960143000489

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2005.03692

Huang, K. J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

Jones, C. R., & Bergen, B. (2024). Does word knowledge account for the effect of world knowledge on pronoun interpretation?. *Language and Cognition*, 1–32. https://doi.org/10.1017/langcog.2024.2

Kim, H., & Shin, G-H. (2021). Effects of long-term language use experience in sentence processing: Evidence from Korean. *Journal of Psycholinguistic Research, 50*, 523–541. https://doi.org/10.1007/s10936-020-09737-0

Kim, J. B., Sells, P., & Yang, J. (2005). Deep Processing of Honorification Phenomena in a Typed Feature Structure Grammar. In *ACL Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Kim-Renaud, Y.-K. (2001). Change in Korean honorifics reflecting social change. In T. E. McAuley (ed.), *Language change in East Asia* (pp. 27–46). London: Curzon.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

Kwon, N., & Lee, Y. (2024). When grammaticality is intentionally violated: Inanimate honorification as a politeness strategy. *Journal of Pragmatics*, 232. 167–181. https://doi.org/10.1016/j.pragma.2024.08.012

Kwon, N., & Mun, S. (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* (pp. 26–37). Association for Computational Linguistics.

Kwon, N., & Shin, G-H. (2025). Overgeneralization of Korean subject honorification by English-speaking learners of Korean. *Bilingualism: Language and Cognition*. https://doi.org/10.1017/S1366728925100813

Kwon, N., & Sturt, P. (2016). Attraction effects in honorific agreement in Korean. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00093

Kwon, N., & Sturt, P. (2019). Proximity and same case marking do not increase attraction effects in comprehension: Evidence from eye-tracking experiments in Korean. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01854

Kwon, N., & Sturt, P. (2024). When social hierarchy matters grammatically: Investigation of the processing of honorifics in Korean. *Cognition*, 251, 105912. https://doi.org/10.1016/j.cognition.2024.105912

Lee, S. & Wang, S. (2023). Do language models know how to be polite?, *Society for Computation in Linguistics 6*(1), 375–378. doi: https://doi.org/10.7275/8621-5w02

Lenth, R. (2025). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.11.1).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Marvin, R., & Linzen, T. (2019). Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics, 2*(1), 373–374. https://doi.org/10.48550/arXiv.1808.09031

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.

Oh, B. D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence, 5*, 777963. https://doi.org/10.3389/frai.2022.777963

Oh, B. D., & Schuler, W. (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. https://doi.org/10.1162/tacl_a_00548

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34*, 739–773. https://doi.org/10.1006/jmla.1995.1033

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2023). The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 20827–20905).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Shin, G-H., & Mun, S. (2025). Modelling child comprehension: A case of suffixal passive construction in Korean. *Computer Speech & Language*, *90*, 101701. https://doi.org/10.1016/j.csl.2024.101701

Sohn, H. M. (1999). *The Korean language*. Cambridge: Cambridge University Press.

Song, S., Choe, J. W., & Oh, E. (2019). An empirical study of honorific mismatches in Korean. *Language Sciences*, 75, 47–71. https://doi.org/10.1016/j.langsci.2019.101238

Sravanthi, S., Doshi, M., Tankala, P., Murthy, R., Dabre, R., & Bhattacharyya, P. (2024). PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 12075–12097). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.719

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford University Press.

Sun, K., & Wang, R. (2025). Computational Sentence-Level Metrics of Reading Speed and Its Ramifications for Sentence Comprehension. *Cognitive Science*, *49*(7), e70092. https://doi.org/10.1111/cogs.70092

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*(3), 285–318. https://doi.org/10.1006/jmla.1994.101

de Varda, A. G., Marelli, M., & Amenta, S. (2024). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods, 56*(5), 5190–5213. https://doi.org/10.3758/s13428-023-02261-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates, Inc.

Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1737–1743) Cognitive Science Society. https://doi.org/10.48550/arXiv.2007.06761

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics, 7*, 625–641. https://doi.org/10.1162/tacl_a_00290

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 211–221). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1809.00042

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The Linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1052

Yoo, M. H., Kim, J., & Song, S. (2025). Multilingual capabilities of GPT: A study of structural ambiguity. *PLoS One*, *20*(7), e0326943. https://doi.org/10.1371/journal.pone.0326943

Zhou, X., Chen, D., Cahyawijaya, S., Duan, X., & Cai, Z. (2025, January). Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6866–6888). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.459/