UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Syntactic Prominence and Pragmatic Bias in Turkish Subject Anaphora: Humans and Large Language Models

Permalink

https://escholarship.org/uc/item/7d46z2b9

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Onal, Esra

Kübler, Sandra

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Syntactic Prominence and Pragmatic Bias in Turkish Subject Anaphora: Humans and Large Language Models

Esra Önal (esraonal@iu.edu) & Sandra Kübler (skuebler@iu.edu)
Department of Linguistics, Indiana University

Abstract

We examined whether large language models (LLMs) responded to syntactic prominence of discourse entities and pragmatic biases in Turkish subject anaphora resolution similar to human judgments. Using an offline comprehension task with native speakers, we showed that null and overt pronouns responds to syntactic prominence and pragmatic bias. We then evaluated several autoregressive LLMs on the same materials. While GPT-4o correlated with human responses, LLaMA-4 more closely approximated the interaction between syntactic prominence and pragmatic biases observed in human data, raising questions about the relationship between model performance, scale, and human fit. We also found that all models mostly differed from humans in response variability, with model responses tending to be more deterministic. Finally, the results indicated partial but limited alignment between human and model anaphora resolution in Turkish.

Keywords: Anaphora resolution, Turkish, Large Language Models

Introduction

The study of anaphoric expressions has long been a central topic in theoretical linguistics, psycholinguistics, natural language processing, and understanding (NLP/NLU) (Ariel, 1988; Grosz et al., 1983; Hobbs, 1978; Mitkov, 2014; Winograd, 1972), given their role in tracking and retrieving discourse entities during real-time language production and comprehension. Their resolution requires the integration of multiple pieces of structural and contextual information. As a result, anaphora resolution has become an important benchmark for evaluating artificial language systems (Levesque et al., 2012).

In this study, we examine how closely LLM interpretations align with human responses to referential ambiguity under pragmatic manipulations, where there is no single correct resolution. We focus on a canonical pro-drop language, Turkish, which allows null subjects and thus provides an informative test case. In Turkish, null and overt subject pronouns are not interchangeable but are associated with distinct discourse functions (Enç, 1986; Taylan, 1986; Turan, 1996), and both production and interpretation depend on the cognitive status of discourse entities as well as on the form(Ariel, 1988; Gundel et al., 1993).

Our study addresses three main questions. First, what factors define human interpretation of null and overt pronouns in intra-sentential contexts, and how do syntactic prominence and pragmatic bias jointly shape these interpretations? Second, how do LLMs integrate syntactic and pragmatic information? And what information do they appear to rely on most strongly? Third, to what extent do LLM and human interpretations align across different dimensions of comparison?

To answer these questions, we combine behavioral data from native Turkish speakers with analyses of multiple autoregressive transformer-based language models. Human participants completed an offline comprehension task in which anaphoric expressions (null vs. overt) and pragmatic bias (neutral, subject-biased, object-biased) were manipulated. We then evaluated model behavior on the same materials by analyzing generated responses. By comparing humans and models across these dimensions, we aim to clarify which aspects of anaphora resolution LLMs capture well and where their behavior diverges from human judgements.

Anaphora Resolution

Establishing discourse coherence requires linking successive utterances, using anaphoric expressions such as pronouns to refer to previously mentioned discourse entities, *antecedents* (Grosz et al., 1983). Inappropriate anaphoric choices disrupt coherence and increase processing cost, reflecting a mismatch between an anaphoric expression and the accessibility of its intended referent (Almor, 1999; Gordon et al., 1993).

Languages differ in the anaphoric expressions they make available. In non–pro-drop languages such as English, subjects must be overtly realized, allowing speakers to choose between pronouns and repeated names (e.g., *Mary/she came home*). In contrast, many pro-drop languages (e.g., Turkish, Italian, Spanish) permit subjects to be phonologically unrealized, giving rise to zero or null pronouns (e.g., Turkish: *(Mary/o) eve geldi* '(Mary/she) came home'). Although these forms may be truth-conditionally equivalent, their distribution and interpretation are constrained by discourse-pragmatic, cognitive, and grammatical factors (Ariel, 1988; Carminati, 2002; Filiaci et al., 2014; Gundel et al., 1993).

A central concept in accounts of anaphora resolution is *prominence*, often understood as the relative accessibility or activation of discourse entities in memory (Ariel, 1988; Grosz et al., 1983; Gundel et al., 1993). Prominence is influenced by multiple cues, including grammatical role (e.g., subjecthood), topicality, linear position, and discourse relations (Carminati, 2002; Gordon et al., 1993; Grosz et al., 1983; Özge & Evcen, 2020). Many approaches to anaphora resolution converge on

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

the idea that more prominent referents are preferentially expressed using less informative anaphoric expressions (Ariel, 1988; Gundel et al., 1993), consistent with Gricean expectations about informativeness (Almor, 1999; Grice, 1975).

Under Accessibility Theory (Ariel, 1988), anaphoric expressions are ordered along a scale from highly informative to highly reduced, with reduced forms signaling that the referent is relatively more accessible in the mental representation of discourse entities. While the relative ordering of anaphoric expressions appears broadly stable across languages, the functional range and spacing of individual forms are language-specific. Filiaci et al. (2014) argues that Spanish overt pronouns align more closely with null pronouns than Italian overt pronouns, which shows a clearer functional distinction from the null pronoun.

Psycholinguistic evidence shows that mismatches between anaphoric expressions and prominence incur processing costs. In syntax-oriented languages such as English and Italian, grammatical subjects are typically treated as the most prominent antecedents, giving rise to the subject bias in null pronoun interpretation (Carminati, 2002; Gordon et al., 1993). Violations of these expectations incur costs such as the Repeated Name Penalty and, in pro-drop languages, the Overt Pronoun Penalty (Lezama & Almor, 2011). In intra-sentential anaphora, (Carminati, 2002) argues that this creates a characteristic division of labor: null pronouns show a strong bias toward subject antecedents, while overt pronouns tend to favor non-subject antecedents, also known as the Position of Antecedent Strategy (POS); however, Filiaci et al. (2014) shows that the strength of this non-subject bias is language-specific. In the current study, we test this hypothesis in Turkish, comparing human judgments with the behavior of LLMs.

Turkish null and overt subject anaphora

Turkish is a canonical pro-drop language that allows null subjects. Although verbal morphology encodes person and number, the use of null versus overt subject pronouns is not optional, but constrained by discourse-pragmatic factors (Enç, 1986; Öztürk, 2002; Taylan, 1986). Accordingly, there is broad agreement that null and overt subject pronouns in Turkish are not freely interchangeable. Null subjects are strongly associated with topic continuity and highly accessible antecedents, whereas overt pronouns tend to appear in marked discourse contexts, such as contrast or topic shift (see Table 1).

Corpus-based work shows that topic shifts—especially shifts to previously mentioned non-subjects—are often realized by fuller forms such as overt pronouns or repeated noun phrases (Turan, 1996). Experimental studies further indicate that anaphora resolution in Turkish interacts with verb semantics and discourse relations. For example, coherence relations linked to psychological and emotion predicates such as causal relations introduced by *because* generate expectations that bias interpretation toward either the subject or the object, sometimes overriding prominence-based expectations (Özge & Evcen, 2020; Özge et al., 2025). However, Özge and Evcen (2020) also show that while null pronouns generally maintain a subject-oriented bias, overt pronouns allow interpretation to track semantic expectations. This asymmetry indicates a robust syntactic bias for null, but not for overt pronouns.

Table 1: Topic continuity and shift in Turkish (Ø = null pronoun; adapted from (Öztürk, 2002; Taylan, 1986)).

Context:
Ben$_i$ ev-e gel-di-m.
I home-DAT come-PST-1SG ‘I came home.’

a. Topic continuity
Ø$_i$ biraz kitap oku-du-m.
PRO a.bit book read-PST-1SG ‘I did some reading.’

b. Topic shift / contrast
Ama John$_j$/o$_j$ gel-me-di.
but John/he come.NEG-PST.3SG ‘But John/he didn’t come.’

Anaphora resolution in LLMs

Anaphora resolution poses a challenging task for LLMs, as it requires the integration of syntactic, semantic, and pragmatic information, along with world knowledge to achieve human-like comprehension. Using surprisal as a measure of model expectations, Davis and van Schijndel (2020) show that LSTM language models are largely insensitive to discourse structure, failing to reflect implicit causality biases in pronoun prediction. In contrast, the transformer-based GPT-2 XL exhibits human-like sensitivity to implicit causality: pronouns consistent with the discourse-preferred antecedent receive lower surprisal than those referring to less preferred antecedents.

Michaelov and Bergen (2022) show that for Italian zero anaphora, autoregressive transformers predict human reading times more reliably than masked transformer language models (e.g., BERT), due to greater sensitivity to word order and positional information, and that performance depends on training data: multilingual exposure to pro-drop languages helps, whereas fine-tuning on English, which lacks null subjects, does not provide the relevant distributional evidence. Zhu et al. (2025) further show that LLMs’ apparent discourse sensitivity is largely driven by lexical cues rather than by abstract discourse-state representations, resulting in divergences from human judgments under lexical variation.

Although it does not involve a direct comparison with human response data, Oğuz et al. (2024) examine a linguistic phenomenon in Turkish known as indexical shift, where the first-person indexical *ben* “I” in finite embedded clauses can be interpreted either as the actual speaker or as the attitude holder (the main subject). They find that LLMs select shifted or non-shifted readings primarily based on contextual cues, while largely ignoring clause type, which in Turkish grammatically determines whether indexical shift is licensed. As a result, large multilingual models such as GPT-4 often allow shifted interpretations even in nominalized clauses, where in-

Table 2: Stimuli illustrating anaphoric expression × pragmatic bias.

Pragmatic Bias	Null vs. Overt Pronoun
Neutral	*Elif Ayşe-yi aradığı-nda,* Elif Ayşe-ACC call-WHEN *(o) yemek yi-yor-du.* (she) food eat-PROG-PAST.3RD "When Elif called Ayşe, (she) was eating."
Subject	*Elif Ayşe-yi herkes-in önünde* Elif Ayşe-ACC everyone-GEN in-front-of *küçük düşürdüğü-nde,* humiliate-WHEN *(o) özür dile-di.* (she) apology make-PAST.3RD "When Elif humiliated Ayşe in front of everyone, (she) apologized."
Object	*Elif Ayşe-yi herkes-in önünde* Elif Ayşe-ACC everyone-GEN in-front-of *küçük düşürdüğü-nde,* humiliate-WHEN *(o) ağla-dı.* (she) cry-PAST.3RD "When Elif humiliated Ayşe in front of everyone, (she) cried."

dexical shift is syntactically impossible. This suggests that models override syntactic constraints when contextual evidence favors a particular reading.

Experiments

This study examines the interaction between syntactic and pragmatic biases in the resolution of Turkish subject anaphora within intra-sentential contexts in humans and machines.

Design and Hypotheses

We consider intra-sentential contexts with two competing antecedents that differ in their syntactic role as subject and object. We assume that these entities are hierarchically ordered by prominence, with subjects more prominent than objects, following (Carminati, 2002). Accordingly, in the absence of additional cues, we expect null subjects to preferentially resolve to the subject antecedent, consistent with topic continuity, and overt pronouns to favor non-subject antecedents, consistent with their discourse-marked use and association with topic shift in Turkish (Enç, 1986; Taylan, 1986; Turan, 1996).

In pragmatically manipulated contexts, however, the interpretation of the anaphor—regardless of its form—is biased toward a specific discourse entity, either the subject or the object. Given that this study uses an offline comprehension task, two alternative hypotheses are considered. Under a full pragmatic override account, discourse bias alone determines anaphora resolution, and both null and overt pronouns follow the same interpretation in biased contexts. Under a persistent syntactic bias account, form-based preferences continue to influence final responses, such that null and overt pronouns differ in how strongly they align with pragmatic bias. These hypotheses can be distinguished by whether responses converge across pronoun types or remain different. If the latter holds, this would suggest that grammatical subjecthood remains a strong source of prominence in Turkish anaphora resolution, even when pragmatic cues are available.

Comparing human and model responses allows us to assess whether language models replicate the balance between syntactic prominence and pragmatic bias observed in humans. For language models, we test whether anaphora resolution is guided primarily by pragmatic bias or whether sensitivity to syntactic prominence is also reflected in model outputs. Under a pragmatic-dominant account, models are expected to follow contextual bias uniformly, showing no significant differences between null and overt pronouns in biased contexts. Under an interaction account, models are expected to reflect form-based differences, such that syntactic prominence modulates how strongly model outputs align with pragmatic bias.

Experiment 1: Humans

Methods Twenty-three native Turkish speakers (9 men, 14 women; aged 18–35) participated in the study by completing an online questionnaire via Google Surveys. Participants were recruited via online social media platforms and screened to exclude individuals with formal training in linguistics. The sample size was chosen to allow the detection of relatively large effects in anaphora resolution across experimental conditions.

A total of 93 experimental sentences were constructed, crossing two factors: anaphoric expression (null vs. overt pronoun) and pragmatic bias (no bias, subject bias, object bias). Each item consisted of a *when/while* adverbial subordinate clause followed by a main clause, following the design of Experiments 1 and 2 in Carminati (2002). The subordinate clause introduced two antecedents—one in subject position and one in object position—while the main clause contained the pronoun (null or overt). Representative stimuli are shown in Table 2. For each item, participants were asked a comprehension question (e.g., "Who was crying?") and indicated which of the two discourse referents they interpreted the pronoun as referring to, choosing between the subject, the object, or a third option labeled *başkası* ("someone else").

Responses were analyzed at the trial level using mixed-effects logistic regression models. Separate binary models were fitted for each response category (subject, object, and someone else), predicting the probability of selecting that re-

sponse as a function of anaphoric expression (null vs. overt), pragmatic bias (neutral, subject bias, object bias), and their interaction. All models included random intercepts for participants and stimuli to account for repeated measures.

Results The analyses revealed a robust syntactic bias distinguishing null and overt pronouns (see Figure 1, Neutral). Null pronouns showed a strong preference for subject antecedents, whereas overt pronouns disfavored subject interpretations, instead shifting reference toward non-subject antecedents. Relative to null pronouns, overt pronouns were associated with fewer subject responses ($\beta = -2.948$, SE = 0.199, $p < .001$) and more object ($\beta = 1.794$, SE = 0.170, $p < .001$) and *other* responses ($\beta = 3.888$, SE = 0.431, $p < .001$).

Pragmatic discourse biases significantly affected interpretation but differed for null and overt pronouns. Subject-biased contexts increased subject interpretations and decreased object interpretations for both pronoun types (all ps $< .01$), with no significant interaction (all $|\beta| < 1$, all ps $> .05$). In contrast, object-biased contexts showed an interaction with the anaphor form: although object bias increased object interpretations ($\beta = 1.963$, SE = 0.238, $p < .001$), this effect was significantly weaker for overt than for null pronouns ($\beta = -1.502$, SE = 0.347, $p < .001$).

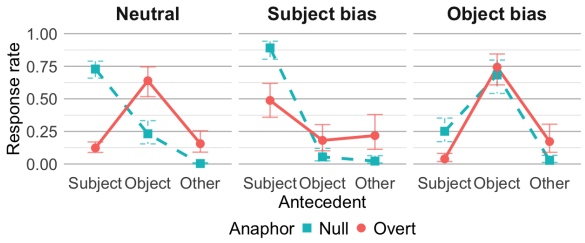

Figure 1: Human response rates by bias and form (95% CIs).

Discussion Our results revealed a significant effect of anaphoric expression on antecedent interpretation, supporting the view that null and overt pronouns in Turkish are not freely interchangeable in use (Enç, 1986; Taylan, 1986; Turan, 1996). In the intra-sentential contexts tested here, null pronouns exhibited a strong preference for subject antecedents, whereas overt pronouns were more often associated with non-subject antecedents, showing a categorical asymmetry (null → subject; overt → non-subject). This was also consistent with accessibility-driven accounts of anaphora resolution, in which reduced forms preferentially target highly accessible or prominent antecedents while fuller forms signal reduced accessibility (Ariel, 1988, 1991; Gundel et al., 1993). This relationship between anaphoric expression and syntactic role further supported the view that grammatical subjecthood constituted a robust source of discourse prominence (Gordon et al., 1993; Grosz et al., 1983; Turan, 1996), especially in intra-sentential anaphora resolution (Carminati, 2002).

Although null pronouns initially favored syntactically prominent antecedents in neutral contexts, pragmatic biases were able to re-rank discourse entities, increasing the relative prominence of the object and allowing non-canonical interpretations in the final resolution. This showed that the syntactic bias of null pronouns was modulated by pragmatic cues and thus could extend to non-subject antecedents successfully. This differed from the findings reported by Özge and Evcen (2020), where null subjects remained subject-oriented across implicit causality contexts, likely reflecting differences in syntactic and discourse structure. Under subject-biased contexts, however, overt pronouns did not show a corresponding increase in subject interpretations; instead, we observed a slight increase in *other* responses, despite an overall bias toward the subject. One possible explanation for this draws on the Gricean principle of Quantity and informativeness: when discourse bias further increases the prominence of an already highly accessible subject, the use of an overt pronoun—being more informative than a null pronoun—may be perceived as over-informative. As a result, participants may have treated the overt pronoun as signaling a departure from the maximally prominent antecedent, increasing uncertainty or weakened commitment to the subject interpretation.

Table 3: Example prompt and model response.

Prompt	*When Hasan phoned Ayşe, (she) was eating.* Who was eating? **Choose one of the options:** Hasan; Ayşe; someone else. **Answer:**
Response	In the given sentence, Ayşe was eating. Therefore, the correct answer is 'Ayşe'.

Experiment 2: Large Language Models

Methods We conducted computational analyses using autoregressive transformer-based LLMs, especially models from the LLaMA-Instruct family (Touvron et al., 2023) and GPT-4o (Hurst et al., 2024). We evaluated both multilingual and Turkish fine-tuned LLaMA models. The multilingual models included LLaMA-4 and LLaMA-3. The Turkish fine-tuned model, TR-LLaMA-3, was included for direct comparison with the base LLaMA-3 model[1]. All LLaMA models were accessed via Hugging Face. Because the models differ in size, training, and optimization, we retained the default settings for each model rather than enforcing identical decoding parameters across models. Specifically, we used temperature = 0.7 and top_p = 0.95 for LLaMA-4, and temperature = 0.6 and top_p = 0.9 for LLaMA-3 and TR-LLaMA-3.

The goal was to examine whether these models resolve anaphora in a similar way to humans in Turkish, and to what extent they rely on the syntactic prominence versus pragmatic biases. The experimental design closely paralleled the behavioral studies: model prompts were constructed to mirror the human experimental questions and were treated as a proxy

[1]LLaMA-4 = `meta-llama/Llama-4-Scout-17B-16E-Instruct`; LLaMA-3 = `meta-llama/Meta-Llama-3-8B-Instruct`; TR-LLaMA-3 = `ytu-ce-cosmos/Turkish-Llama-8b-Instruct-v0.1`.

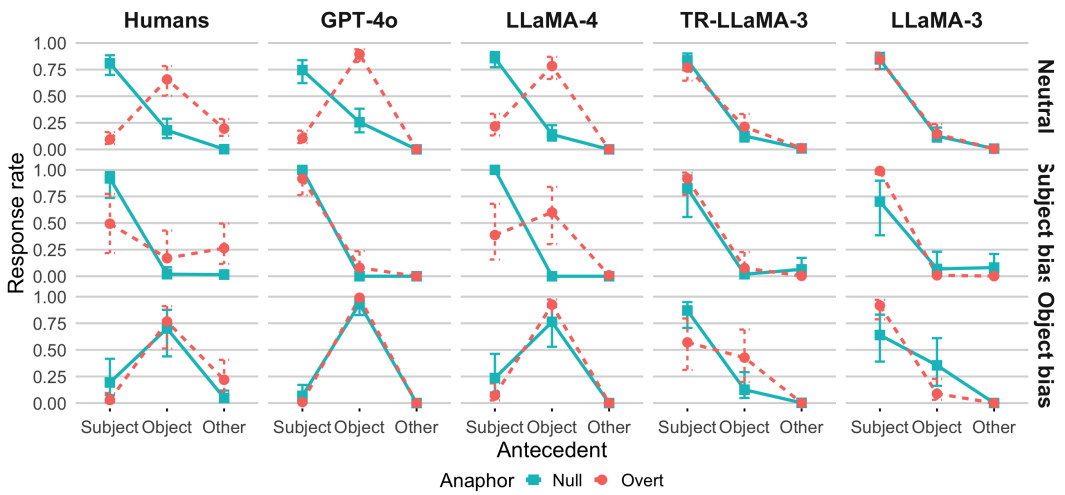

Figure 2: Human and model response rates by bias and anaphoric expression (95% CIs).

for participant responses. To encourage focused outputs, we appended the instruction to each prompt, as highlighted in Table 3. The models were allowed to freely generate responses of up to 128 tokens.

Each model was sampled 100 times per stimulus. Because generated outputs were often lengthy and included explanations (see example in Table 3), we manually coded the final choice expressed in each response. Some responses were excluded under two conditions: when the model conflated the subject and object names into a single entity (e.g., "Ayşe Fatma"), likely due to the SOV word order of the stimuli, or when the model failed to select either option. Exclusion rates ranged from 0% for LLaMA-4 to 4.09% for TR-LLaMA-3 (GPT-4o: 0.04%, LLaMA-3: 1.04%).

Human–model differences were evaluated using regression-based analyses and Pearson's correlation coefficient (ρ). Responses were aggregated across sampling iterations into sentence-level counts for each interpretation type (subject, object, other) and analyzed as weighted observations, with humans as the reference level. Sentence-level dependence arising from repeated sampling was accounted for using mixed-effects models with random intercepts for sentences. The primary analyses consisted of binomial mixed-effects regression models with a logit link, contrasting subject versus non-subject interpretations (object + other) as a function of anaphoric expression (null vs. overt), bias condition (neutral, subject bias, object bias), and model, including their full interaction. In addition to the full model including humans and all LLMs, separate binomial mixed-effects models were fit for each individual model to assess model-specific effects of anaphoric expression and bias. Statistical significance of fixed effects was assessed using Wald z-tests, and ρ was used to quantify human–model similarity based on sentence-level response distributions. Finally, response variability was quantified using entropy (Shannon, 1948) over sentence-level subject response probabilities.

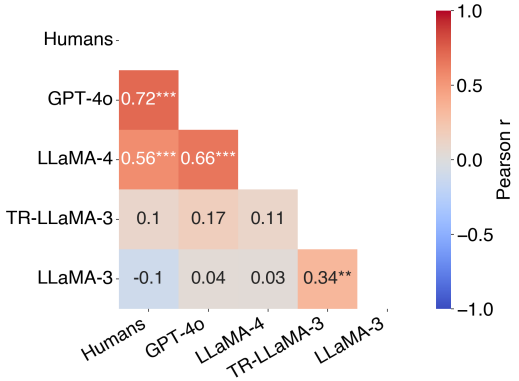

Figure 3: Correlation of subject response rates (* p < 0.05, ** p < 0.01, *** p < 0.001; Bonferroni-corrected).

Results Analyses conducted separately for each language model indicated differences across models. **GPT-4o** showed a strong baseline preference for subject interpretations with null anaphors (intercept: $z = 7.69$, $p < .001$), which was significantly reduced for overt pronouns ($z = -9.89$, $p < .001$). Significant interactions between anaphoric expression and pragmatic bias (subject bias: $z = 3.58$, $p < .001$; object bias: $z = 5.13$, $p < .001$; Fig. 4) indicated that discourse bias can override syntactic preferences. Moreover, relative to the human baseline, GPT-4o exhibited greater sensitivity to pragmatic bias (subject bias: $z = 4.76$, $p < .001$; object bias: $z = -3.11$, $p = .002$). However, model comparisons revealed a significant divergence from humans in the anaphor–bias interaction only for subject bias ($z = -2.47$, $p = .013$), but not for object bias ($z = -0.84$, $p = .40$), reflecting a reduced overt–null contrast under subject bias in GPT-4o.

LLaMA-4 was similar to GPT-4o in its sensitivity to syntactic prominence, but exhibited a strong interaction only in the *object-bias* condition ($z = 4.07$, $p < .001$). In contrast, subject bias did not significantly modulate the effect of

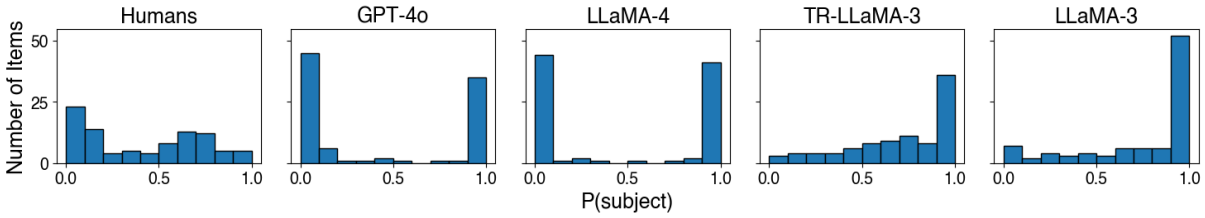

Figure 4: Item-level variation in humans and LLMs.

Table 4: Mean entropy of subject response probabilities.

System	Mean Entropy
Humans	0.663
GPT-4o	0.139
LLaMA-4	0.080
TR-LLaMA-3	0.528
LLaMA-3	0.341

anaphoric expression ($z = -0.09$, $p = .93$), indicating that subject bias failed to override the strong syntactic preferences associated with anaphor type. Consistent with the human baseline, the overt–null contrast under object bias was preserved in LLaMA-4, with no reliable evidence that the magnitude of the overt–null contrast in LLaMA-4 differed from that observed in humans under either subject bias ($z = -0.78$, $p = .44$) or object bias ($z = -0.71$, $p = .48$).

Neither **LLaMA-3** nor the Turkish fine-tuned **TR-LLaMA-3** showed reliable main effects of anaphoric expression or pragmatic bias, and both exhibited an overall subject preference across bias conditions for both pronoun types.

Correlation results (Figure 3) showed that alignment with human responses increased with model scale.[2] LLaMA-4 deviated least from the human baseline in regression analyses, while GPT-4o showed the strongest item-level correlation with human responses. Distributional analyses showed that human subject responses clustered around intermediate probabilities, reflecting human variability, whereas model responses concentrated near extreme values, indicating more deterministic behavior (see Figure 4). Entropy analyses indicated that, unlike humans, all models except TR-LLaMA-3 produced highly deterministic item-level distributions (Table 4).

Discussion Overall, the results indicated that models, except GPT-4o, exhibited reduced sensitivity to either syntactic biases associated with anaphoric expression or pragmatic biases. In contrast, GPT-4o exhibited strong sensitivity to pragmatic cues, often overriding syntactic preferences, consistent with a pragmatic-dominant resolution strategy that differed from humans, similar to findings reported by Oğuz et al. (2024), who showed that large language models often prioritize contextual cues over syntactic constraints in reference resolution.

Regression analyses and item-level correlations (ρ) highlighted potentially different dimensions of human–model similarity, differentiating how models arrived at responses from whether they arrived at the same response. While GPT-4o aligned more closely with human outcomes, LLaMA-4 more closely captured the interaction between syntactic and pragmatic biases in humans. This dissociation suggests that models can reproduce human responses while differing in underlying processes by which linguistic cues are combined, although the extent and interpretation of these differences require further investigation / experimental evidence. This account is consistent with recent work showing that increases in model scale and success in next-word prediction do not necessarily entail closer alignment with human linguistic behavior (Oh & Linzen, 2025). Larger models can provide a worse fit to human reading times and neuroimaging measures (Lin & Schuler, 2025), as they tend to be "less surprised" than humans (Oh & Schuler, 2023). Oh and Linzen (2025) attribute this mismatch to differences between human and model memory constraints, which may suggest a reversed U-shaped relationship between model scale and human fit.

The response probability distributions and entropy measures indicated that LLMs produced more deterministic responses than humans, with probability mass concentrated near extreme values and lower entropy at the item level[3]. In contrast, human responses exhibited greater variability across participants. Our results suggested that models may resemble human behavior along some dimensions (e.g., outcome alignment or interaction structure), while differing in the extent of response variability observed in humans.

We also found that model behavior might be driven by lexical and form-specific factors. For LLaMA-4, certain verb pairings (e.g., *humiliate–apologize*) under subject-biased contexts caused near-categorical object interpretations with overt pronouns, while others remained highly sensitive to subject bias. This item-specific variability suggested limited cross-item generalization, consistent with evidence that LLM anaphora resolution can rely on lexical cues rather than abstract discourse representations (Zhu et al., 2025). This differed somewhat from human behavior, where overt pronouns still showed an overall tendency toward subject interpretations, even as syntactic prominence shifted responses toward non-subject interpretations relative to null pronouns.

[2] An analysis of object responses produced comparable results.

[3] To control for sampling effects, we repeated the analyses using prob. derived from model output logits, with comparable results.

Acknowledgments

We thank the Indiana University Department of Linguistics and the College of Arts and Sciences for financial support related to this research, including support through a College of Arts and Sciences Travel Award. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. Portions of this work were presented at the Cognitive Science Program and the Department of Linguistics at Indiana University, and we thank members of these communities for valuable feedback and discussion.

References

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological review*, *106*(4), 748.

Ariel, M. (1988). Referring and accessibility. *Journal of linguistics*, *24*(1), 65–87.

Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*(5), 443–463. https://doi.org/https://doi.org/10.1016/0378-2166(91)90136-L

Carminati, M. N. (2002, February). *The Processing of Italian Subject Pronouns* [Doctoral dissertation, University of Massachusetts Amherst].

Davis, F., & van Schijndel, M. (2020, November). Discourse structure interacts with reference but not syntax in neural language models. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 396–407). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.32

Enç, M. (1986). Topic Switching and Pronominal Subject in Turkish. In D. I. Slobin & K. Zimmer (Eds.), *Studies in Turkish Linguistics* (pp. 195–209). John Benjamins.

Filiaci, F., Sorace, A., & Carreiras, M. (2014). Anaphoric biases of null and overt subjects in Italian and Spanish: A cross-linguistic comparison. *Language, Cognition and Neuroscience*, *29*(7), 825–843.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, Names, and the Centering of Attention in Discourse. *Cognitive Science*, *17*(3), 311–347.

Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, *3*, 43–58.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *21st Annual Meeting of the Association for Computational Linguistics*, 44–50. https://doi.org/10.3115/981311.981320

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, *44*(4), 311–338. https://doi.org/https://doi.org/10.1016/0024-3841(78)90006-2

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. *KR*, *2012*(13th), 3.

Lezama, C. G., & Almor, A. (2011). Repeated Names, Overt Pronouns, and Null Pronouns in Spanish. *Language and Cognitive Processes*, *26*(3), 437–454.

Lin, Y.-C., & Schuler, W. (2025). Surprisal from larger transformer-based language models predicts fmri data more poorly. *arXiv preprint arXiv:2506.11338*.

Michaelov, J. A., & Bergen, B. K. (2022, October). Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns? In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 1–14). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.1/

Mitkov, R. (2014). *Anaphora resolution*. Routledge.

Oğuz, M., Ciftci, Y., & Bakman, Y. F. (2024, August). Do LLMs recognize me, when I is not me: Assessment of LLMs understanding of Turkish indexical pronouns in indexical shift contexts. In D. Ataman, M. O. Derin, S. Ivanova, A. Köksal, J. Sälevä, & D. Zeyrek (Eds.), *Proceedings of the first workshop on natural language processing for turkic languages (sigturk 2024)* (pp. 53–61). Association for Computational Linguistics. https://aclanthology.org/2024.sigturk-1.5/

Oh, B.-D., & Linzen, T. (2025). To model human linguistic prediction, make llms less superhuman. *arXiv preprint arXiv:2510.05141*.

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.

Özge, D., & Evcen, E. (2020). Referential form, word order and emotional valence in turkish pronoun resolution in physical contact events. In D. Zeyrek & U. Özge (Eds.), *The view from turkish* (pp. 165–186). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110686654-007

Özge, D., Evcen, E., & Hartshorne, J. K. (2025). Referential form, word order, and implicit causality in turkish emotion verbs. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Öztürk, B. (2002). Turkish as a non-pro-drop language. In E. E. Taylan (Ed.), *The Verb in Turkish* (pp. 239–260, Vol. 44). John Benjamins Publishing Company.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Taylan, E. E. (1986). Pronominal versus zero representation of anaphora in turkish. In D. I. Slobin & K. Zimmer (Eds.),

Studies in turkish linguistics (pp. 209–232). John Benjamins Publishing Company. https://doi.org/doi:10.1075/tsl.8.12tay

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Turan, Ü. D. (1996). *Null vs. overt subjects in turkish discourse: A centering analysis* [Doctoral dissertation, University of Pennsylvania].

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*(1), 1–191.

Zhu, X., Zhou, Z., Charlow, S., & Frank, R. (2025). Meaning beyond truth conditions: Evaluating discourse level understanding via anaphora accessibility. *arXiv preprint arXiv:2502.14119*.