

From LLM Errors to Linguistic Theory: Human Evaluation of Japanese Particle Predictions

Yurie Hara¹, Hoi Man Kevin Li², and John S. Y. Lee²

¹Research Faculty of Media and Communication, Hokkaido University; hara@imc.hokudai.ac.jp

²Department of Linguistics and Translation, City University of Hong Kong

Japanese sentence-final particles (SFPs) play a crucial role in discourse management and interpersonal communication. Their interpretation depends on subtle pragmatic factors such as speaker intentions, assumptions about shared knowledge, and interactional goals (Katagiri, 2007; McCready & Nomoto, 2024). Although Japanese SFPs have been extensively studied in linguistics, relatively little work has examined their automatic prediction, with the notable exception of Mandokoro et al. (2024). A strict comparison between an LLM prediction and the particle attested in a corpus assumes that the latter is uniquely determined. Yet written transcripts typically omit prosody, gaze, gesture, intended addressee, and aspects of shared background and speaker stance. A mismatch may therefore reflect either a genuine model failure or a context that licenses more than one pragmatically plausible SFP.

RESEARCH QUESTIONS AND GOAL: We ask (i) how much SFP choice can be recovered from textual discourse context, (ii) whether LLM predictions that differ from corpus-attested particles are rejected by native speakers of Japanese, and (iii) whether LLM–corpus mismatches are associated with human uncertainty. We combine a large-scale corpus evaluation with two completed human experiments: a cloze task eliciting particle choices from textual contexts, and a seven-point acceptability-judgment task comparing corpus-attested (Gold) and LLM-predicted particles. This comparison among corpus production, LLM prediction, cloze choice, and acceptability treats model–human divergence as evidence about pragmatic underspecification, without assuming that every divergence has the same source.

CORPUS-BASED ANALYSIS OF LLM PREDICTIONS: The Nagoya University Conversation Corpus (Fujimura et al., 2012) contains 129 dialogues, segmented into 129,221 sentences. Of these, 27,788 contain a discourse SFP and served as target sentences. We tested DeepSeek under three prompting methods: a zero-shot prompt with the target sentence alone (*no context*); the target plus the preceding ten dialogue turns (*with context*); and retrieval-augmented generation (RAG), which additionally supplied five similar turns retrieved from the rest of the corpus. Because the model sometimes produced no SFP, the main exact-match comparison used the 26,426 sentences for which all three methods generated one.

Exact-match accuracy increased from 45.86% with no context to 52.79% with dialogue context and 54.40% with RAG (Table 1). Context thus improved recovery of the attested particle, but even the best-performing method achieved only moderate exact-match accuracy. The dominant residual pattern was overgeneration of *yo*. Under RAG, *yo* occurred 8,497 times versus 5,323 corpus occurrences; frequent mismatches included Gold *ne* → LLM *yo* (1,276 cases) and Gold *yone* → LLM *yo* (894). In (1)–(2), the corpus speaker uses *ne*, plausibly presenting the approximate exchange rate as shared and inviting alignment, whereas RAG chooses *yo*, plausibly presenting it as new information. Crucially, the contrast need not correspond to a difference in truth conditions: the propositional content is held constant while the particle changes how the speaker positions that content relative to the interlocutors’ discourse states.

The choice between *ne* and *yo* can therefore depend on assumptions about epistemic access and alignment that are only partially represented in a written transcript. Exact-match prediction of SFPs consequently tests not only linguistic competence but also recovery of latent discourse states from incomplete contextual evidence. Such contrasts motivate human evaluation because the transcript alone may not determine the interlocutors’ epistemic states. The human experiments therefore examine both the acceptability of the differing alternative and the extent to which readers independently recover the corpus particle.

HUMAN CLOZE EXPERIMENT: The cloze experiment involved 50 native speakers of Japanese and 100 target items (30 Gold=LLM; 70 Gold≠LLM). A mixed-effects logistic regression revealed a strong association between LLM and human performance: participants were significantly more likely to select the corpus-attested particle when the LLM also predicted it ($\beta = 1.61$, $SE = 0.41$, $z = 3.92$, $p < .001$; OR

= 5.00, 95% CI [2.23, 11.17]). Model-estimated Gold-selection probabilities were 59.3% for Gold=LLM items versus 22.6% for Gold $\neq$ LLM items (Figure 1a).

LLM–corpus mismatches were also associated with greater dispersion in human responses across alternative particles. Mean response entropy was higher for Gold $\neq$ LLM items than for Gold=LLM items (1.69 vs. 1.31 bits; mean difference = 0.38 bits, Welch’s $t(45.69) = 2.41$, $p = .020$; Figure 1b). This difference was also supported by a rank-based test and a participant/item bootstrap analysis.

Together, these results show that LLM–corpus mismatches are associated with two distinct signatures of human uncertainty: lower recovery of the corpus-attested particle and greater dispersion across alternative SFPs. Crucially, the entropy effect shows that the lower Gold-selection rate does not simply reflect a systematic shift from Gold to a single competing particle; rather, human responses are distributed more broadly across alternatives when the LLM diverges from the corpus. Thus, LLM–corpus mismatch systematically tracks the degree to which textual discourse context constrains human SFP choice.

HUMAN ACCEPTABILITY EXPERIMENT: In the acceptability experiment, 50 native speakers of Japanese judged 50 target items on a seven-point scale. Each discourse context appeared with either its corpus-attested particle (Gold) or the differing LLM-predicted particle. The experiment also included natural and anomalous fillers, providing a baseline against which the target contrast could be interpreted. As Figure 2 shows, anomalous fillers received markedly lower ratings than natural fillers and target items, whereas the corpus-attested and LLM-predicted target conditions showed strikingly similar rating distributions. Thus, the absence of a large difference between the two target conditions cannot be attributed simply to participants’ failure to distinguish more from less acceptable particle uses.

Because responses were ordinal, the cumulative link mixed model (CLMM) was primary. It found no significant difference between the corpus-attested and LLM-predicted conditions ($\beta = -0.149$, $p = .159$). This nonsignificant result does not itself establish equivalence. Separately, the raw difference was small (approximately 0.19 points on the seven-point scale, with the LLM-predicted condition slightly lower), and two one-sided tests (TOST) established equivalence within the predefined ± 0.5 -point bound. Thus, while participants readily penalized anomalous filler items, the differing LLM-predicted particles were, on average, nearly as acceptable as the corpus-attested particles.

INTEGRATED INTERPRETATION AND LINGUISTIC IMPLICATIONS: The two human tasks probe complementary aspects of SFP choice. The cloze task measures which particle readers independently recover from a transcript, whereas the judgment task asks whether a particular particle remains acceptable once it is supplied. A context may therefore fail to lead readers reliably to the corpus-attested particle while still licensing both the corpus and LLM alternatives. The relevant notion of “error” therefore depends on the evaluation criterion: failure to reproduce a speaker’s corpus-attested choice is not equivalent to producing a pragmatically infelicitous particle.

This dissociation between recoverability and acceptability suggests that SFP choice is gradient at the level of discourse appropriateness: textual context may favor one particle without rendering alternatives infelicitous. Corpus production therefore provides evidence about speaker choice, but not necessarily a categorical boundary between felicitous and infelicitous SFPs.

Taken together, the two human experiments show that many apparent LLM errors cannot straightforwardly be treated as failures to identify the uniquely appropriate SFP. LLM–corpus mismatches are concentrated in contexts where humans themselves are substantially less likely to recover the corpus-attested particle, while the LLM alternatives remain comparably acceptable in an independent judgment task. Thus, strict exact-match accuracy risks conflating genuine model failures with cases in which the textual context does not uniquely constrain SFP choice. The present results do not isolate which pragmatic information gives rise to this indeterminacy: shared knowledge, speaker intention and stance, prosody, gaze, and multi-party addressee structure remain plausible contributors. Comparing corpus production, model predictions, and human responses therefore provides a diagnostic for how strongly SFP choice is constrained by transcript-visible discourse and supplies empirical targets for further pragmatic analysis.

Table 1. DeepSeek exact-match accuracy

Prompting method	No context	With context	RAG
Exact-match accuracy	45.86%	52.79%	54.40%

26,426 common-output sentences.

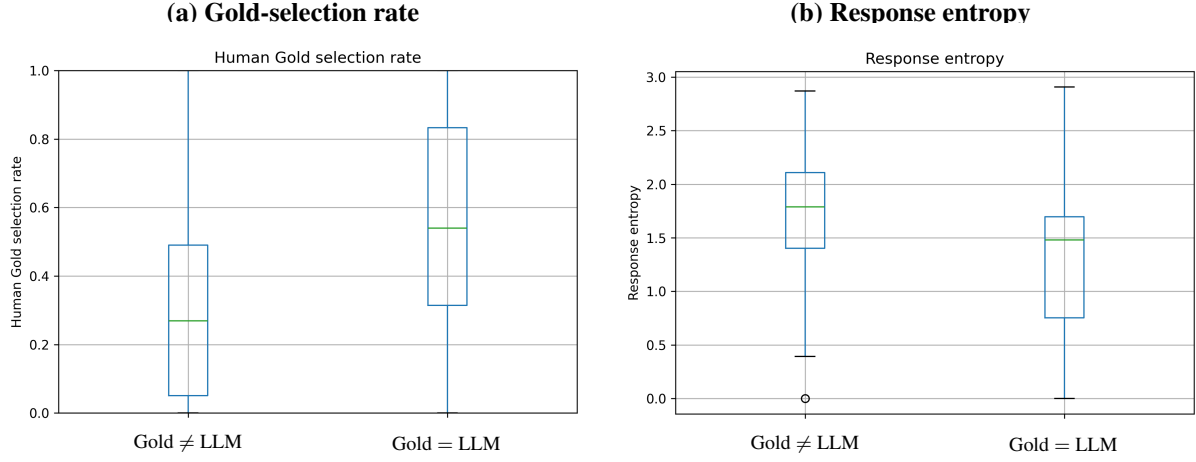


Figure 1. Human cloze performance as a function of LLM agreement with the corpus: (a) Gold-selection rate and (b) response entropy. Gold $\neq$ LLM: 70 items; Gold=LLM: 30 items.

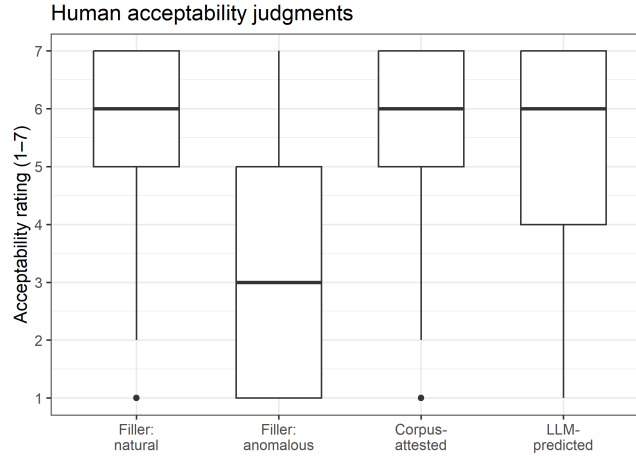


Figure 2. Human acceptability judgments across filler and target conditions.

- (1) F128 : 7 0 パー引き? 本が? (びっくりしました) (70% off? The books? [surprised])
F023 : ほいでねえ。 (And then...)
F128 : そいで買ったんだ。 (So you bought it.)
F023 : あきに、あきに、私がバービーちゃんとか、それからドール、60年代の人形のコレクションの写真
つきのハードカバーが (ふーん) これって、裏返したらUS\$30ドルしてるんだね。
(For example, a hardcover book with photos of Barbie dolls and other doll collections from the 1960s—when you turned it
over, the list price was US\$30.)
F023 : USドルで。 (In US dollars.)
F107 : それがたぶん1000円ぐらいだった【ね】。 (That came to about 1,000 yen-ne.)
- (2) *Sore-ga tabun 1000-yen gurai dat-ta ne.*
that-NOM probably 1000-yen approximately COP-PAST SFP
‘That was probably about 1,000 yen, wasn’t it?’

SELECTED REFERENCES: Fujimura et al. 2012. Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. Katagiri. 2007. Dialogue functions of Japanese sentence-final particles ‘Yo’ and ‘Ne’. Mandokoro et al. 2024. Understanding of utterance meaning by Subjective-BERT: An analysis of internal representation with particular reference to self-attention chain. McCready & Nomoto. 2024. *Discourse Particles in Asian Languages. Volume I: East Asia*. Routledge Studies in Linguistics.