



LLMs GO AROUND RATING SENTENCES VERY POSITIVELY WITHOUT ATTENTION TO AFFECTIVE MEANING: AN EXPERIMENTAL COMPARISON OF LLM AND HUMAN SENSITIVITY TO CONSTRUCTIONAL SEMANTIC PROSODY

VICTORIA KRUMPHOLZ *

Abstract *Semantic prosody describes the affective meanings or connotations of a linguistic element, which are commonly understood in terms of positive and negative alignment. Under the view of Construction Grammar, not only individual words but larger linguistic structures may carry semantic and pragmatic meaning, including positive or negative semantic prosody. The present pilot study takes this approach as a point of departure to investigate LLMs' sensitivity to subtle pragmatic meanings in the form of semantic prosody. 4 state-of-the-art LLMs as well as a human participant group were tasked with rating the connotation and pleasantness of the go-around-Ving construction said to carry negative semantic prosody as well as two syntactically parallel constructions with neutral/positive semantic prosody. Results showed that while LLMs exhibit some sensitivity to constructional semantic prosody, their rating behavior differed significantly from humans when pooled into one group. Compared to humans, LLMs gave higher median and mean connotation and pleasantness ratings while also assigning less neutral ratings across constructions. Although not generalizable due to the exploratory character of this study, the results demonstrate that subtle pragmatic phenomena like semantic prosody represent a promising research area when it comes to discerning LLM language abilities and how they compare to humans'.*

Keywords: semantic prosody, Construction Grammar, LLM language abilities, pragmatics, large language models

1. INTRODUCTION

Only two months after its launch in November 2022, ChatGPT¹ rose to 100 million total users and set the record for being 'the fastest growing consumer application in history' (Hu 2023). Over three years later, generative AI applications are more relevant than ever, with popular functions ranging from coding to text generation and machine translation. Unsurprisingly, scholarly interest in the large language models (LLMs) commercial AI systems are built upon has grown alongside ChatGPT's userbase. Various opinions have emerged in the field of linguistics about the use LLMs for the field of linguistics: some scholars advocate for

* Freie Universität Berlin

¹ OpenAI. 2022. "ChatGPT." <https://chatgpt.com/>.

the use of LLMs in developing linguistic theory (cf. Milli re 2024), while others warn against LLMs' limitations in language comprehension and production when compared to humans (cf. Carchidi 2024). Despite differing perspectives, the central question asked by scholars remains the same: How do LLM language abilities, including both comprehension and production, compare to those of humans?

The possible ways to address this question are numerous, spanning all formal linguistic domains from morphology to semantics as well as adjacent areas such as language acquisition, biases and creative language use. Simultaneously, experimental comparisons of LLM and human language are shaped by the theoretical frameworks they are based on, leading to a near infinite number of potential studies needed to comprehensively answer the question above. Nonetheless, each empirical investigation adds to our understanding of LLM language and how it compares to the natural language(s) used by humans.

The present study seeks to make such a contribution in the form of an exploratory pilot study offering a first look at a phenomenon from the domain of pragmatics: semantic prosody (cf. Bednarek 2008), here used to refer to subtle affective meanings of linguistic expressions which may have an effect on the reading of the surrounding linguistic context. Adopting a Construction Grammar (CxG) perspective, this study compares human and LLM affective ratings of the *go-around-Ving* construction alleged to have negative semantic prosody (Jensen 2024) to two syntactically parallel constructions (*go-on-Ving* and *dance-around-Ving*) which have not been associated with negative affect. Due to its exploratory nature, this work does not test directional hypotheses about group differences. Rather, its aim lies in examining whether human and LLM judgements vary systematically across constructions with different semantic prosody. Furthermore, results may reveal whether human participants perceive *go-around-Ving* as carrying any negative affective meaning.

Section 2 provides an overview of the theoretical background and relates this study to previous research. While Section 3 details the experimental design, Section 4 presents the statistical results, which are subsequently related to the research question and methodological limitations in Section 5. Section 6 concludes the study, offering an outlook on further research situated at the interface of LLM and human language and pragmatics.

2. THEORETICAL BACKGROUND

As previously mentioned, the present study takes up the framework of Construction Grammar (cf. Goldberg 1995, Hilpert 2019) as a perspective on natural language systems, adopting the notion that grammatical constructions are triplets of (syntactic and phonetic) form and meaning with open slots for lexical items to be inserted. Under this view, the strict division between lexicon and grammar disappears as meaning is no longer limited to lexemes and words but borne by larger grammatical structures as well. Constructional meaning is often used to explain variation in argument structure, where the element inserted into an open constructional slot takes on the construction-inherent argument structure. This forced semantic typeshift caused by the constructional context is also known as coercion and provides an explanation for spontaneous and creative language use (cf. Hilpert 2019, 17).

However, constructional meaning is not limited to the semantic domain. It also extends to pragmatic phenomena like semantic prosody, and pragmatic constructional meaning may influence how the construction-external linguistic context is interpreted. While the term semantic prosody has been used to refer to a number of different concepts including collocational phenomena, this paper follows the definition given by Bednarek (2008), who distinguishes between semantic preference and semantic prosody. While semantic preference refers to collocational patterns exhibited by a linguistic structure (e.g. preferentially collocating with items from a specific semantic class), semantic prosody describes the connotations, i.e. ‘attitudinal and functional meanings’ (Bednarek 2008, 131), of a linguistic item, which are often understood in terms of positive/negative alignment. Combining this notion with the concept of constructional meaning, it seems possible that complex linguistic structures may exhibit positive or negative semantic prosody and carry attitudinal meaning, which then might influence language users’ understanding of elements occurring in the respective construction’s open slot(s) as well as the surrounding context.

An example of this is the *go-around-Ving*² construction, which has been described by lexicographers as carrying a negative connotation (for our purposes: semantic prosody) which does not arise from its structural components, as Jensen (2024, 2) notes. The *Cambridge Learner’s Dictionary*³ exemplifies this by listing the meaning of the British variant *go round* + *gerund* as ‘behaving badly’ as well as ‘to spend your time behaving in the stated way’, suggesting an element of habituality in addition to clear negative semantic prosody (cf. Jensen 2024, 4 for more examples). While the *Oxford Learner’s Dictionary*⁴ does not provide an explicit reference to constructional meaning, the example listed (*It’s unprofessional to go around criticizing your colleagues*) is clearly in line with the negative semantic prosody postulated above. As Jensen (2024) shows in his corpus study comparing the collocational behavior of *go-around-Ving* and *go-around-and-V*, negative semantic prosody converges with semantic preference of negatively connotated verbs in the V2 slot.

Based on the negative semantic prosody of *go-around-Ving* postulated above, it seems plausible that when neutral or positively connotated verbs or phrases are inserted into the constructional V2 slot, they are colored by the negative semantic prosody of *go-around-Ving* such that the whole sentence takes on a negative affective meaning. Consider (1): A native speaker of English might have the intuition that this sentence conveys a subtly negative meaning, although another sentence like (2) might not, showing that a possible negative semantic prosody of (1) must be related to the *go-around-Ving* construction. In addition, Jensen (2024, 4) mentions that *go-around-Ving* expresses unpleasantness which can be experienced by anyone involved in the event described by the full sentence, i.e. the agent, the patient or a witness. Therefore, speakers might also judge (1) to convey a more unpleasant meaning than (2).

² *Ving* represents an open verb slot, filled by a verb in the present participle form. Henceforth, this slot will be referred to as the constructional V2 slot, **with the lemma go representing V1**.

³ go around. 2025. In *Cambridge Learner’s Dictionary*. “go around.”
<https://dictionary.cambridge.org/dictionary/learner-english/go-around>.

⁴ *Oxford Learner’s Dictionary*. “go around (sense 3).”
https://www.oxfordlearnersdictionaries.com/definition/english/go-around_1?q=go+around.

- (1) My mother went around telling all her friends that I had a very nice girlfriend.
- (2) My mother told all her friends that I had a very nice girlfriend.

If it is indeed the case that native English speakers perceive a difference in semantic prosody between (1) and (2), which one might test in an experiment contrasting *go-around-Ving* with parallel constructions (see Section 3.2), this difference is likely subtle and socially learned, forming part of the linguistic domain of pragmatics⁵.

Returning to the issue of how LLMs comprehend and produce language compared to humans, semantic prosody appears to be a promising phenomenon for study. Most recent studies comparing LLM and human language abilities focus on grammatical and semantic phenomena, e.g. language learning mechanisms and plural formation (Sauerland et al. 2025), lexico-syntactic flexibility and generalization over non-prototypical forms (Mortensen et al. 2024) and semantic meaning comprehension (Dentella et al. 2024). Pragmatic factors like world knowledge play a minor role in Zhou et al.’s 2024 examination of three construction types containing *so ... that*, although the study’s main focus lies on semantic meaning. While Combs et al. (2025, 12) compare LLM and human affective ratings of linguistic terms, finding that LLMs tend to assign less neutral ratings despite strong overall correlation with human ratings, their study employs a sociological rather than a linguistic perspective.

As illustrated above, the present study represents a novel inquiry into the pragmatic abilities of LLMs, specifically their sensitivity to affective constructional meaning and its effects on the constructional context. By combining the CxG approach to language with the notion of semantic prosody, the present study attempts to provide an insight into how sensitive humans and 4 state-of-the-art LLMs are to affective constructional meaning, contributing to our understanding on how LLM language abilities differ from or converge with those of humans overall.

3. METHODOLOGY

3.1 Groups

To compare LLMs’ sensitivity to constructional semantic prosody to that of humans, four state-of-the-art models were prompted: GPT-5⁶, Gemini 2.5 Flash⁷, DeepSeek V3⁸ and Llama-3.3-70B⁹. Models were selected based on free accessibility and the availability of a web-based interface for querying. Out of the four models prompted in this experiment, the former two represent proprietary, closed models while DeepSeek V3 and Llama-3.3-70B are open-source

⁵ This may seem somewhat counterintuitive given the term semantic prosody. Both Jensen (2024: 4) and Bednarek (2008: 131) note this discrepancy, and ultimately, it seems justified to classify semantic prosody as a pragmatic phenomenon as it refers to affective and inferential meaning.

⁶ OpenAI. 2025. “GPT-5 (large language model).” <https://chatgpt.com/>.

⁷ Google DeepMind. 2025. “Gemini 2.5 Flash (large language model).” <https://gemini.google.com/>.

⁸ DeepSeek. 2025. “DeepSeek V3 (large language model).” <https://www.deepseek.com/>.

⁹ Meta. 2025. “Llama-3.3-70B (large language model).” <https://poe.com/Llama-3.3-70B>.

models with architectures available to the public.

Human language data was collected from five native English speakers who completed an online questionnaire on [soscisurvey.de](https://www.soscisurvey.de)¹⁰ (Leiner 2025), a German platform for scientific data collection in accords with the GDPR. All data was collected between October 1–7, 2025.

3.2 Materials

To test whether human language users perceive the *go-around-Ving* construction to carry negative semantic prosody and/or convey notions of unpleasantness, the construction under investigation was contrasted with two syntactically parallel constructions differing with regards to the motion verb and particle respectively. As it is unclear whether the alleged negative semantic prosody of *go-around-Ving* is related to the particle *around* or the motion verb *go*, both possibilities were accounted for by comparing *go-around-Ving* to the construction *go-on-Ving*, as well as *dance-around-Ving*. As *dance* likely carries a positive affective meaning in comparison to *go*, this allows for the investigation of an effect of the verb’s positive semantic prosody on the event encoded in a full sentence.

Four sentences containing the string *go around Ving* were extracted from COCA (Davies 2008–) to ensure occurrence in natural speech. The query used is presented in (3).

(3) GO around _v?g

An important criterion in sentence selection was that the VP filling the constructional V2 slot must not have negative semantic prosody on its own, as this could influence participants’ rating of the sentence. Instead, only sentences containing VPs with neutral or positive semantic prosody were chosen, so that any negativity effect on ratings could be traced back to construction-inherent semantic prosody.¹¹ Other criteria specified that VP must be formally compatible with all constructions tested (*go-around-Ving*, *go-on-Ving*, *dance-around-Ving*), sentences must not contain negation as this, once again, could yield more negative ratings, and that sentences must sound natural. The four sentences chosen are presented in (4a–d), with *go* and *around* being replaced by *dance* and *on* to form the relevant parallel sentences for filler constructions, resulting in a set of 12 total stimuli.

(4)

- (a) Why does everyone *go/dance around/on acting* like it's true?
- (b) We *went/danced around/on humming*.
- (c) We *went/danced around/on talking* about utopian visions.
- (d) My mother was *going/dancing around/on telling* all her friends that I had a very nice

¹⁰ Leiner, Dominik J. 2025. *SoSci Survey*. <https://www.soscisurvey.de/>.

¹¹ Due to time constraints, the identification of neutral or positive semantic prosody in VPs was conducted solely by the author. Ideally, statistical inter-rater reliability measures should be computed for at least two independent raters, to ensure robustness of judgements.

girlfriend.

Due to not all sentences meeting the criteria above, minor adjustments had to be made for sentences (4c) and (4d): For (4c), negation was removed and the tense was changed from present to past for a more natural-sounding sentence while (4d) is a fragment of a longer sentence.

Stimuli were presented to human and LLM participants in the same order and with the same instructions. Participants were asked to consider the stimuli sentence which was presented as an utterance made by some speaker A. They were then asked to answer two questions on a 5-point Likert scale, with scale labels listed explicitly for each rating. The questions are presented in Table 1, along with explanatory notes to maximize instruction intelligibility.

Table 1 Questions and explanatory notes

No	Question	Scale
1	How would you rate the event described with regards to its connotation? Explanatory notes: Event described: e.g. go/dance around/on acting like it's true Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning	1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)
2	How would a person affected by the event described feel? Explanatory notes: Person affected: Could be the agent (e.g. everyone) or someone who happens to witness the event described	1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

Question 1 targets semantic prosody while question 2 focuses on the unpleasantness notion ascribed to *go-around-Ving* in various dictionaries (see Section 2). Connotation was chosen as a near synonym for semantic prosody in question 1 to ensure participants' familiarity with the term. The exact wording of the explanatory notes was adjusted on a case-by-case basis to fit the stimulus sentence presented. Notes were presented for each stimulus to account for the fact that context windows (i.e. previous input utilized in producing output) are difficult to establish in LLMs, especially in closed models like GPT-5.

Before being presented with the target sentences shown in (4a–d), participants were given a brief introduction to the task in which they were asked to answer spontaneously based on their intuition. Afterwards, they had to rate two example stimuli and one practice item before proceeding to the experiment. As practice stimuli, three sentences with strong positive/negative semantic prosody were chosen to familiarize participants with the experiment. After completing the practice, the experimental stimuli were presented to participants individually, with sentence and construction types varying randomly, although the same construction was never presented twice in a row.

For LLM prompting, an additional instruction line was added asking the model to provide only the numerical rating as well as the corresponding scale label. This adjustment was made after one model (Llama-3.3-70B) had already been prompted, in reaction to a failure to follow instructions by Gemini 2.5 Flash (see Section 5). The instruction was therefore only presented to 3 out of 4 of the LLMs tested here.

GPT-5 and Gemini 2.5 Flash were prompted via the ChatGPT web app¹² and the Gemini web app¹³ while both DeepSeek V3 and Llama-3.3-70B were prompted via Poe¹⁴ where a multitude of models are available freely. Due to the model size of DeepSeek V3, the daily prompting limit was reached on Poe while conducting the experiment, leading to an administering of experimental prompts across multiple sessions. To account for an effect of user adaptation on experimental ratings, new accounts were created on each web app, and apps were opened in an anonymous window in Google Chrome.

As stated above, prompts were presented to human participants in form of a questionnaire on *soscisurvey.de*. Participants had to give informed consent before being allowed to proceed with the study. In addition to the instructions and stimuli described above, the online questionnaire included a short demographic survey to ensure that all participants self-identified as native English speakers.

All experimental materials, including a full version of the script used for LLM prompting as well as the online questionnaire, can be found in the Appendix.

3.3 Statistical Analysis

Data analyses were conducted in R 4.5.1 (R Core Team 2025) using RStudio, with scripts generated and adjusted via GPT-5¹⁵. No AI-generated texts were reproduced directly in this work. For data handling and visualization, the tidyverse suite (Wickham et al. 2019) was utilized. Specialized packages used for fitting models were *brglm2* (Kosmidis 2025) and *ordinal* (Christensen 2023).

Ratings obtained by humans and LLMs were combined into one data set, which then underwent removal of practice items, standardization of predicator labels and separation into two subsets by question (connotation/pleasantness). LLMs were treated individually unless stated otherwise, to avoid generalizations across LLMs with differing architectures where possible. Descriptive statistics included median and interquartile range (IQR) to respect the ordinality of the data. In addition, mean and SD were calculated and used for visualization of rating patterns by *Group* and *Construction*. Although the treatment of Likert-scale data as approximately interval-scaled is common practice, the ordinality of the data should be kept in mind when interpreting the results. Bar plots showing rating distributions by *Construction* and *Group* were also produced. To account for a possible effect of *SentenceType*, mean ratings ($\pm$ 95% confidence intervals) by *SentenceType* and *Group* were computed and visualized in a dot

¹² chatgpt.com.

¹³ gemini.google.com.

¹⁴ Quora, Inc. 2025. "Poe." <https://poe.com/>.

¹⁵ OpenAI. 2025. GPT-5 (large language model, accessed via ChatGPT Plus). <https://chatgpt.com/>.

plot.

For the inferential analysis, *Human* and *go-around* were set as baselines in the *Group* and *Construction* categories respectively. Due to sparse cell counts for individual LLMs across the five-point scale, the rating variable was first recoded as binary (1–2 vs. 3–5) to fit bias-reduced binary logistic regression models (method = ‘brglmFit’, Kosmidis 2025) examining *Group* and *Construction* effects for connotation and pleasantness data. To retain the information encoded in the full five-point scale, ordinal regression models, specifically cumulative link models with a logit link (method = ‘clm’, Christensen 2023) testing for effects of *Group2* and *Construction* were also fitted. For this purpose, LLMs were pooled into one group (*Group2* variable) to increase cell counts. All models were compared and tested for significance using likelihood-ratio tests (LRTs) and drop1 analyses. The modelling approaches described here should be understood as exploratory, testing for systematic rating differences by group and construction rather than fully directional hypotheses.

A full version of the workflow can be found in the Appendix.

4. RESULTS

4.1 Descriptive Analysis

Table 2 contains the full summary of *Rating* median, IQR, mean and SD for each *Group* and *Construction* combination for both connotation and pleasantness data. On the 5-point Likert scale employed in the experiment, higher numbers indicate more positive/pleasant ratings, with 3 representing neutrality. As noted above, means and SDs are reported as additional descriptive measures and should be interpreted alongside medians and IQRs, keeping in mind the ordinal nature of the data.

Table 2 Descriptive statistics summary (lowest medians and means per construction/question highlighted in bold)

Question	Group	Construction	N	Median	IQR	Mean	SD
Connotation	Human	go-around	20	3	1.25	2.95	1.05
Pleasantness	Human	go-around	20	3	1.25	3.05	0.94
Connotation	Human	go-on	20	3	0.5	3.1	0.91
Pleasantness	Human	go-on	20	3	0.25	3.1	0.79
Connotation	Human	dance-around	20	4	1	3.6	0.94
Pleasantness	Human	dance-around	20	4	1	3.6	0.88
Connotation	GPT-5	go-around	4	4	0.75	3.75	1.26
Pleasantness	GPT-5	go-around	4	4	0.75	3.75	1.26
Connotation	GPT-5	go-on	4	4	0.75	3.75	1.26
Pleasantness	GPT-5	go-on	4	4	0.75	3.75	1.26

Connotation	GPT-5	dance-around	4	5	0.75	4.25	1.50
Pleasantness	GPT-5	dance-around	4	5	0.75	4.25	1.50
Connotation	Llama-3.3-70B	go-around	4	4.5	1.5	4	1.41
Pleasantness	Llama-3.3-70B	go-around	4	4.5	1.5	4	1.41
Connotation	Llama-3.3-70B	go-on	4	4	2.25	3.75	1.50
Pleasantness	Llama-3.3-70B	go-on	4	4	2.25	3.75	1.50
Connotation	Llama-3.3-70B	dance-around	4	4.5	1.5	4	1.41
Pleasantness	Llama-3.3-70B	dance-around	4	4.5	1.5	4	1.41
Connotation	Gemini 2.5 Flash	go-around	4	4	0.5	3.5	1.00
Pleasantness	Gemini 2.5 Flash	go-around	4	4	0.5	3.5	1.00
Connotation	Gemini 2.5 Flash	go-on	4	4	0.5	3.5	1.00
Pleasantness	Gemini 2.5 Flash	go-on	4	4	0.5	3.5	1.00
Connotation	Gemini 2.5 Flash	dance-around	4	4.5	1.5	4	1.41
Pleasantness	Gemini 2.5 Flash	dance-around	4	4.5	1.5	4	1.41
Connotation	DeepSeek V3	go-around	4	4	0.5	3.5	1.00
Pleasantness	DeepSeek V3	go-around	4	4	0.5	3.5	1.00
Connotation	DeepSeek V3	go-on	4	3.5	1.25	3.25	0.96
Pleasantness	DeepSeek V3	go-on	4	3.5	1.25	3.25	0.96
Connotation	DeepSeek V3	dance-around	4	4	0.75	3.75	1.26
Pleasantness	DeepSeek V3	dance-around	4	4	0.75	3.75	1.26

The lowest medians by construction printed in bold show that humans exhibit the lowest median ratings overall – only in one case, namely *dance-around* (DeepSeek V3), does any LLM median rating match the human one. The same trend is continued by the mean ratings: the *Human* group consistently exhibited the lowest mean rating across constructions. Differences between human and LLM mean ratings range from 0.15–0.65 points (*dance-around*) to 0.55–1.05 points (*go-around*) for connotation and from 0.15–0.65 points (*dance-around*) to 0.45–0.95 (*go-around*) for pleasantness.

The general inter-constructional comparison trend reveals that medians and means for *go-around* and *go-on* tend to be slightly lower than for *dance-around* across groups. Connotation and pleasantness medians are identical across groups and mean ratings differ between questions only in the human *go-around* condition.

A visualization of rating proportions by *Construction* and *Group* is provided in Fig. 1a (connotation) and 1b (pleasantness).

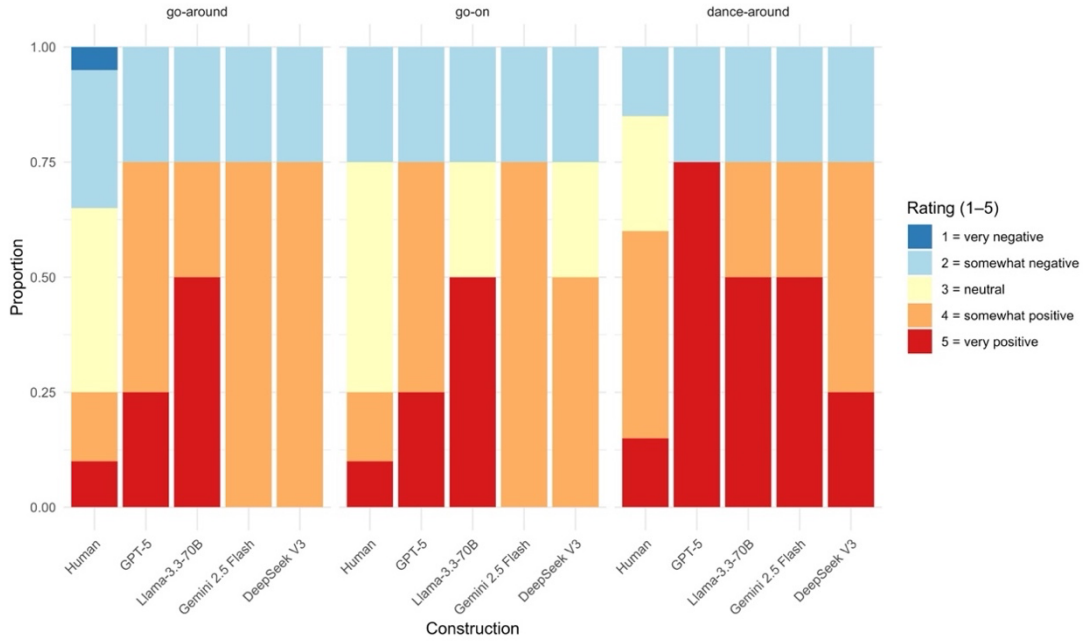


Figure 1a Distribution of connotation ratings by Construction and Group

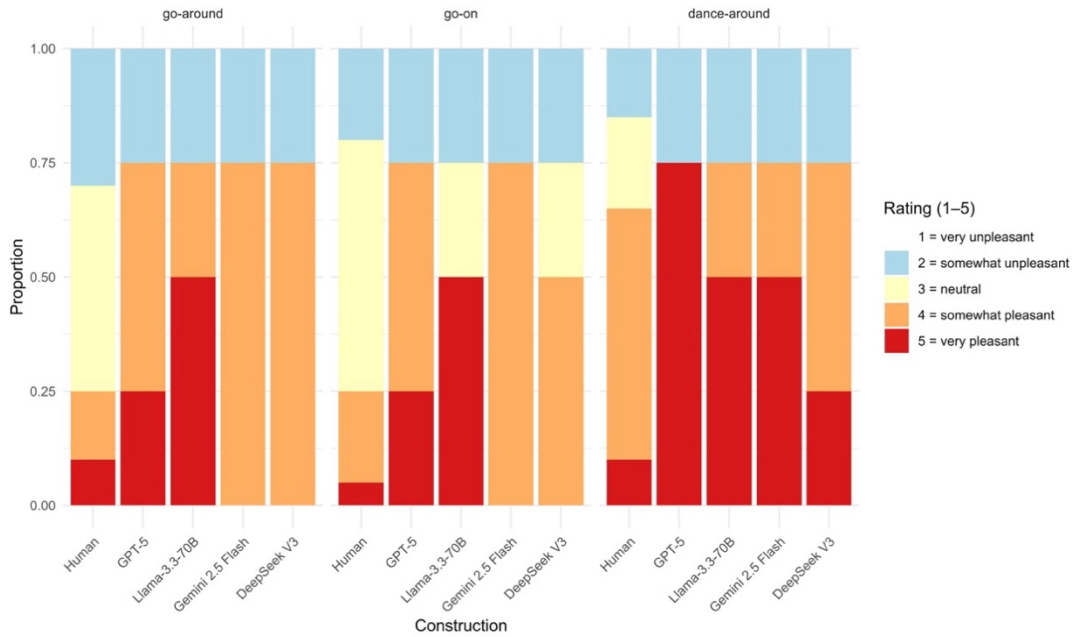


Figure 2b Distribution of pleasantness ratings by Construction and Group

A look at the rating distributions for connotation reveals that while the proportion of somewhat negative (2) ratings remains relatively steady across Construction and Group conditions, GPT-5 and Llama-3.3-70B are more likely than humans to give the highest rating very positive (5) across constructions. Gemini 2.5 Flash and DeepSeek V3 also exhibit higher proportions of this rating than humans for dance-around but did not rate go-around and go-on

as very positive (5) at all. The proportion of somewhat positive (4) ratings also tends to be higher in LLMs for go-around and go-on, but not for dance-around. Furthermore, humans show a higher proportion of neutral (3) ratings across all constructions compared to LLMs, with go-on being the only construction rated neutrally by two LLMs (Llama-3.3-70B and DeepSeek V3). Finally, humans are the only group where a very negative (1) rating was recorded, namely for the go-around construction. Overall, under the assumption that 1–2 represent negative ratings while 3–5 represent non-negative ratings, the differences between humans and LLM ratings are minor, with a slightly higher proportion of negative human ratings for go-around and a lower proportion of negative human ratings for dance-around.

The patterns emerging for pleasantness rating distributions (Fig. 1b) are identical to the connotation rating patterns described for LLMs, which rated connotation and pleasantness the same for each stimulus. Human ratings differ slightly between questions with regards to the go-on construction, which shows a marginally higher proportion of neutral (3) ratings for pleasantness compared to connotation. Strikingly, GPT-5 gave exclusively very positive/pleasant (5) and somewhat positive/unpleasant (2) ratings in the dance-around condition.

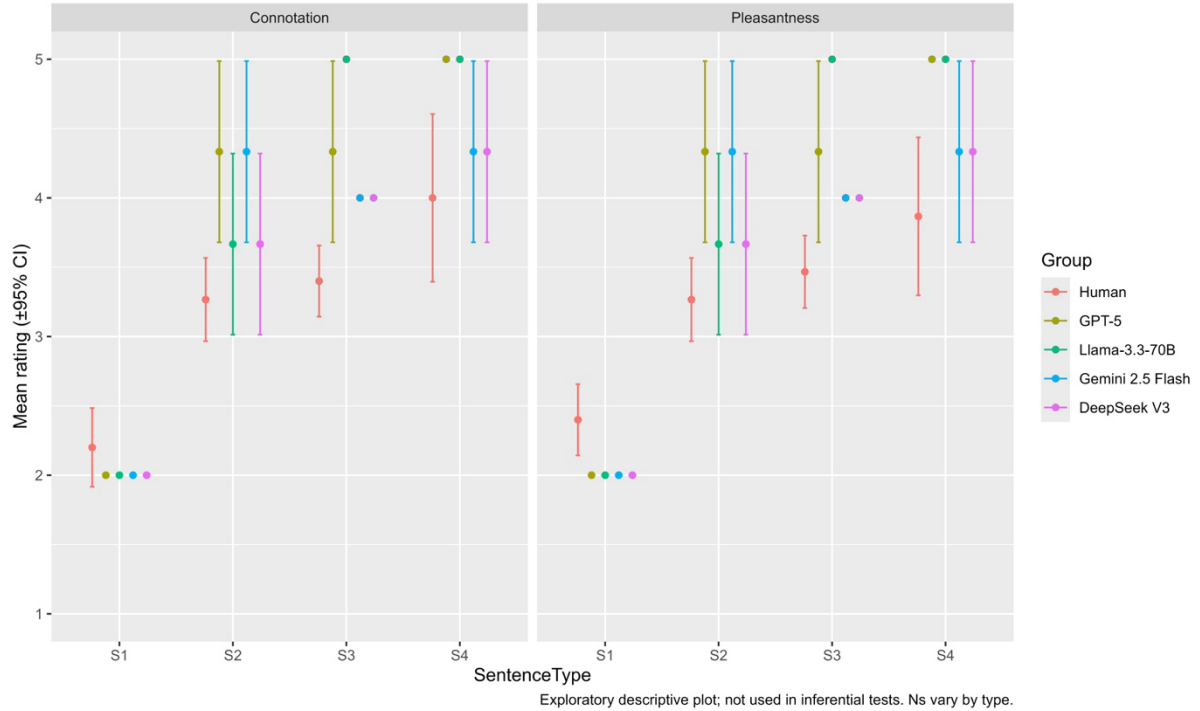


Figure 2 Mean connotation/pleasantness ratings by Group and SentenceType

Finally, the plots displaying mean connotation and pleasantness ratings by SentenceType (with (4a–d) corresponding to S1–S4) and Group (Fig. 2) reveal a striking difference of mean ratings across sentence types for both questions. The mean ratings for S1 (see (4a)) are considerably lower than for other sentences, with S4 (see (4d)) receiving the highest ratings overall. Generally, human mean ratings are slightly higher for S1 than LLM ratings and lower

than LLM ratings for all other sentences.

While the mean rating difference by SentenceType is not revisited in the inferential statistics section due to scope constraints, this finding is briefly discussed in Section 5.

4.2 Inferential Analysis

The inferential analysis examined whether the variables *Group* and *Construction* had a significant effect on connotation and pleasantness ratings, with the null hypothesis stating that neither variable would predict systematic rating differences across conditions. To examine the effect of *Group* and *Construction* variables on the assignment of negative (1–2) vs. non-negative (3–5) ratings while retaining the individual LLM groups, bias-reduced binary logistic models were fit separately for the pleasantness and connotation data. *Human* and *go-around* were set as reference levels. It was established via likelihood-ratio tests that including the *Group* $\times$ *Construction* interaction did not significantly improve model fit ($p_{\text{connotation}} = 1.00$, $p_{\text{pleasantness}} = 1.00$). Subsequent drop1 tests revealed no significant main effects of *Group* ($\chi^2(4) = 0.00$, $p = 1.00$) or *Construction* ($\chi^2(2) = 1.13$, $p = .57$) for the connotation data, with the same pattern holding for the pleasantness data (*Group*: $\chi^2(4) = 0.03$, $p = .99$; *Construction*: $\chi^2(2) = 0.65$, $p = .72$). This was in line with coefficient estimates indicating that no individual LLM diverged significantly from human ratings ($p > .10$) and neither ratings for *go-on* nor *dance-around* differed significantly from *go-around* ($p > .10$) in either data set.

In contrast, when fitting ordinal regression models with *Group2* and *Construction* as predictors after pooling LLMs into one group to mitigate sparse cell counts, drop1 tests showed significant main effects of both *Group2* ($\chi^2(1) = 8.63$, $p = .003^{**}$) and *Construction* ($\chi^2(2) = 6.39$, $p = .04^{*}$) in the connotation data. Coefficient estimates indicated that the pooled LLM group assigned significantly higher ratings than humans ($\beta = 1.07$, $SE = 0.37$, $z = 2.89$, $p = .004^{**}$). Likewise, *dance-around* received significantly higher ratings than *go-around* ($\beta = 0.98$, $SE = 0.44$, $z = 2.23$, $p = .03^{*}$), while no significant rating difference was detected for *go-on* ($p > .10$). Similar patterns emerged in the pleasantness data, where significant main effects were found for *Group2* ($\chi^2(1) = 8.91$, $p = .003^{**}$) and *Construction* ($\chi^2(2) = 6.59$, $p = .04^{*}$). Again, LLMs ($\beta = 1.10$, $SE = 0.38$, $z = 2.93$, $p = .003^{**}$) assigned significantly higher ratings and *dance-around* received significantly higher ratings ($\beta = 0.97$, $SE = 0.44$, $z = 2.18$, $p = .03^{*}$) compared to the respective reference levels, while there was no significant difference found for *go-on* ($p > .10$).

These results suggest that while no significant effects were observed for either *Group* or *Construction* when ratings were pooled into negative (1–2) versus non-negative (3–5) categories to increase cell counts for individual LLMs, taking the full scale into account and focusing on the overall comparison between *Human* and pooled *LLM* groups did reveal significant rating differences, a finding which will be discussed in Section 5.

5. DISCUSSION

The central motivation of this study was to examine whether LLMs' sensitivity to

constructional meaning and subtle pragmatic phenomena like semantic prosody is on par with that of humans. To answer this question, the *go-around-Ving* construction, which is said to have negative semantic prosody and convey unpleasantness, was chosen as an experimental object and contrasted with two syntactically parallel constructions, *go-on-Ving* and *dance-around-Ving*.

In the context of the experimental setup of this study, the research interest described above can be rephrased in the following three questions: (1) Do humans and LLMs rate connotation the same or differently (on a 5-point Likert scale)? (2) Do humans and LLMs rate pleasantness the same or differently? (3) Do human ratings confirm the alleged negative semantic prosody of *go-around-Ving* in comparison with *dance-around-Ving* and *go-on-Ving*?

Table 2 displaying descriptive statistic measures per *Group/Construction* condition distinctly shows that median and mean ratings differ minimally between connotation/pleasantness questions across all groups and constructions. Likewise, significant effect patterns revealed by the ordinal regression models exhibited a high degree of similarity between connotation and pleasantness questions, and an inspection of participants' raw responses confirmed that LLMs always gave the same rating for connotation and pleasantness whereas human ratings differed between questions only in few cases. Given the near-identical rating patterns for connotation and pleasantness, it appears that questions (1) and (2) are best answered together, and that the connotation/semantic prosody and (un)pleasantness conveyed by the constructions tested seem to be closely related.

The group-wise rating proportions (Fig. 1a–b) as well as central tendency measures (Table 2) suggest that human and LLM ratings of the three constructions differ in some respect: human mean ratings were consistently lower than those of LLMs across constructions, with medians for *go-around* and *go-on* also following the same pattern. GPT-5 and Llama-3.3-70B exhibited higher proportions of *very positive* (5) ratings, whereas DeepSeek V3 showed slightly more human-like distributions in the *dance-around* condition. Humans frequently gave *neutral* (3) responses and produced a broader spread of ratings overall, while LLMs rarely assigned neutral ratings and responses never spanned more than three rating categories for each construction. These patterns point towards an avoidance of uninformative (neutral) ratings in LLMs as well as a tendency towards more extreme rating choices compared to humans, which converges with results from Combs et al. (2025, 9, 12).

The bias-reduced binary logistic models did not yield any significant main effects of *Group* and *Construction*, showing that humans were not statistically more likely than individual LLMs to assign explicitly negative ratings (≤ 2) to any construction. The binary recoding of ratings, introduced to handle sparse LLM cell counts, comes at the cost of sacrificing valuable information encoded in the ordinal rating scale, including the neutrality threshold. In contrast, the ordinal regression models, which retained the full 5-point scale, showed highly significant effects of *Group2* on rating behavior ($p < .01$). While pooling all LLMs into one category despite architectural differences represents a methodological limitation, the inferential results align with descriptive patterns and confirm that LLMs do not produce fully human-like ratings, with effect sizes revealing that LLMs gave more positive/pleasant ratings overall.

Although LLM and Human ratings differed significantly across the full scale, between-construction rating differences remained broadly consistent across groups. In addition to the significant main effect of *Group2*, the ordinal regression models revealed a significant main effect of *dance-around*, which received higher ratings than *go-around*. This suggests that LLMs are partially sensitive to constructional semantic prosody, although their ratings appear to be less nuanced than those of humans as illustrated by the lack of neutral ratings assigned by LLMs.

Addressing question (3), human mean ratings tentatively confirm the alleged negative semantic prosody of *go-around-Ving* only in direct comparison to *go-on-Ving* and *dance-around-Ving*. While mean ratings for *go-around-Ving* were marginally lower than for parallel constructions, they approximated neutrality (3) instead of falling categorically in the negative range of the rating scale as previously expected. In addition, the rating difference between *go-around* and *go-on* was not statistically significant across groups, with only *dance-around* showing a robust difference in rating across groups possibly related to a positive semantic prosody of the motion verb *dance*. It appears likely that the effect of *go-around-Ving* on perceived connotation/pleasantness is subtle, arising mostly in direct comparison to other constructions; its detection requires advanced pragmatic reasoning skills. LLMs were not able pick up on this subtle effect in a fully human-like manner, as median and mean ratings of *go-around* and *go-on* (Table 2) suggest.

Another factor which likely influenced results is *SentenceType*. As shown in Fig. 2, *S2–S4* are generally rated somewhat to very positively across groups. While it was not possible to include *SentenceType* as a random effect in the ordinal regression models due to scope constraints of this study, this observation suggests that it is generally possible for sentences containing the *go-around-Ving* construction to be perceived positively overall. This indicates that linguistic context and especially the verb phrase inserted into the V2 slot may counteract the construction’s negative semantic prosody, which becomes apparent only when directly compared to *go-on-Ving* (non-significant) and *dance-around-Ving* (significant).

To sum up, the present study tentatively shows that LLMs did not display fully human-like performance on this rating task. Although models seem to be partly sensitive to semantic prosody of words and phrases, it is not clear whether this extends to very subtle pragmatic constructional meanings like the semantic prosody of *go-around-Ving*, as the statistical patterns above show.

While conducting this study, several methodological limitations emerged in context. First, it is important to note once again that this work represents an exploratory pilot study with the goal of studying LLMs’ sensitivity to phenomena like semantic prosody and pragmatic constructional meaning. Due to sparse counts of (human) participants and stimuli, findings should not be taken as representative. While the patterns observed provide a first insight into how LLMs’ sensitivity to constructional semantic prosody compares to that of humans, they should not be generalized to the broader population. However, they might be utilized in the development of similar experimental tasks as well as the refinement of hypotheses for future studies on semantic prosody.

Regarding the selection of stimuli, the non-assessment of inter-rater reliability represents a methodological limitation to be discussed. The four sentence types selected from the corpus were judged to have neutral or positive semantic prosody by the author; to ensure that stimuli sentence is unbiased and consistent with the experimental aim, future studies should have at least two independent raters evaluate the relevant sentences for semantic prosody, assessing robustness of judgement via statistical measures of inter-rater reliability.

Another methodological limitation of the present study concerns the standardization of Likert-scale data via z-scores, a data preprocessing step meant to reduce inter-participant variability regarding rating scale use. Although standardization represents a measure commonly taken in linguistic studies including judgement data (cf. Schütze & Sprouse 2014), it was not part of the statistical analysis undertaken in this study. Transforming the Likert-scale ratings, while controlling for participants' individual use of the rating scale, moves the focus away from the research question: Do human ratings confirm the alleged negative semantic prosody of *go-around-Ving*? To address this question, the information contained in the raw scale was retained in the statistical analysis; however, it must be noted that due to the small sample size, results may be strongly influenced by inter-participant differences in rating behavior. Z-score standardization represents a valuable robustness measure which should be considered in future studies as an additional step of analysis, as results obtained in this study are interpreted with caution.

Three further limitations should be discussed regarding the inferential statistical analyses conducted: due to sparse data obtained for individual LLMs in this study, fitting ordinal regression models testing for the LLMs' individual performance was not possible. However, increasing data points for LLMs would either require presenting human participants with a large set of stimuli, lowering participant-friendliness, or providing LLMs with additional items, thereby reducing comparability. Nonetheless, treating LLMs as individual groups is desirable, as differing architectures and training data are likely to influence model performance on this task.

Furthermore, the binary recoding of the ordinal 5-point rating scale to remedy sparse cell counts represents another methodological limitation. The significant effects obtained from the ordinal regression models suggest that the binary recoding of the original 5-point scale into negative (1–2) and non-negative (3–5) categories likely resulted in reduced model sensitivity to meaningful variation in the data. As sparse cell counts only became apparent during the analysis, the subsequent binary recoding was an ad-hoc adjustment which does not fully match the original study design. As such, the bias-reduced binary logistic models might have failed to capture subtle semantic prosody effects clustered in different areas of the scale. Additional analyses might test whether a different recoding threshold (e.g. 1–3 vs. 4–5) would yield significant effects; overall, it appears that the threshold chosen here was not fully suitable for detecting significant effects of *Group* and *Construction* predictors in the data, as is confirmed by mean and median ratings overwhelmingly lying in the non-negative (3-5) range.

The last limitation to be addressed here concerns the fitting of the additive ordinal regression models to the pooled-LLM data. Because the binary models showed no significant effect of a *Construction* $\times$ *Group* interaction, the additive ordinal models were tested against intercept-

only models rather than interaction-based ones. This could mean that an interactional effect was missed, which might also be investigated in a more extensive statistical analysis.

Several steps can be taken from here to gain further insight into LLMs' sensitivity to semantic prosody and constructional meaning. While it would be possible to conduct further statistical analyses on the data collected, it appears most promising to build upon the current experimental design and conduct a similar study with a representative population sample, while simultaneously addressing the limitations discussed above. The present study is best viewed a starting point for future investigations into the topic and the results obtained, though tentative, show that pragmatic meaning is a promising research area when it comes to comparing the language abilities of humans and LLMs.

6. CONCLUSION

To add to the growing body on LLM language comprehension abilities in comparison to humans, the present study was designed to test whether LLMs detect and rate subtle pragmatic meanings like semantic prosody in a human-like manner. A construction with alleged negative semantic prosody, *go-around-Ving*, was chosen and contrasted with two constructions believed to have neutral/positive semantic prosody, *go-on-Ving* and *dance-around-Ving*, in a sentence rating task. Statistical analyses revealed that while LLMs exhibit some sensitivity to constructional semantic prosody, their rating behavior differed significantly from humans, which is broadly in line with results from previous studies touching on LLM pragmatic abilities. As a consequence, semantic prosody can be considered a linguistic phenomenon where LLM comprehension does not quite measure up to humans, possibly due to the advanced pragmatic reasoning skills required to detect subtle comparative differences in semantic prosody.

Despite a number of methodological limitations, this study demonstrates the need to examine all domains of language, especially context- and experience-based pragmatic phenomena, when comparing LLM and human language abilities and use. The experimental design of this work as well as the results obtained can act as a point of departure for future studies in LLM-focused linguistics, which might help us get closer to figuring out how not only LLM but human language works.

APPENDIX A. QUESTIONNAIRE AND RATING PROCEDURE (HUMAN GROUP)

This appendix contains instructions given to participants, the two rating questions, the rating scale as well as two representative items. The full questionnaire is available at the following address: [10.5281/zenodo.20642236](https://doi.org/10.5281/zenodo.20642236).

A.1 Participant Instructions

You have decided to participate in the study—welcome! In the following, you will be

presented with **different utterances made by some speaker A**. All utterances contain an event featuring an acting agent. You will then be asked **three different questions**¹⁶ about the event described in the utterance which you are instructed to answer on a scale from 1-5, with scale labels made explicit for each question. You are asked to answer all questions **spontaneously**, based on what comes to your mind first. Below, you find two examples of utterances which might come up as well as explanations on how to answer. Afterwards, you will be asked to complete one training question before moving on to the actual experiment.

Ex. 1

[Utterance]

Consider the following utterance:

A: He has a habit of yelling at his subordinates.

A.2 Questions and Scale

1. How would you rate the event described with regards to its connotation?

Event described: yelling at his subordinates

Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning.

Scale: very negative – somewhat negative – neutral – somewhat positive – very positive

2. How would a person affected by the event described feel?

Person affected: Could be the agent (he), the subordinates or someone who happens to witness the event described.

Scale: very unpleasant – somewhat unpleasant – neutral – somewhat pleasant – very pleasant

A.3 Example Items

[Item 2]

Consider the following utterance:

A: We danced around humming.

[Item 5]

Consider the following utterance:

A: Why does everyone go around acting like it's true?

APPENDIX B. QUESTIONNAIRE AND RATING PROCEDURE (LLM GROUP)

¹⁶ After the fact, it was discovered that both LLM and human questionnaires contained a typo mentioning “three questions” instead of “two”. This may have contributed to LLM failing to follow instructions as expected and should be kept in mind when interpreting the results.

This appendix contains the representative LLM prompts, including instructions and questions, as well as one representative answer per LLM. The full prompting script and chat logs are available at the repository above (see Appendix A).

B.1 LLM Prompt (Instructions)

Introduction and pretraining:

In the following study, you will be presented with **different utterances made by some speaker A**. All utterances contain an event featuring an acting agent.

You will then be asked **three different questions** about the event described in the utterance which you are instructed to rate on a scale from 1-5, with scale labels spelled out explicitly for each question. You are also asked to answer all questions **spontaneously**, based on what comes to your mind first. **Only give the relevant ratings as your answers, explanations are not needed.**

Below, you find two examples of utterances which might come up as well as explanations on how to answer. Afterwards, you will be asked to complete one training question before moving on to the actual experiment.

Ex. 1

[Utterance]

Consider the following utterance:

A: He has a habit of yelling at his subordinates.

[Questions]

1. How would you rate the event described with regards to its connotation?

Event described: yell at his subordinates

Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning

1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)

2. How would a person affected by the event described feel?

Person affected: Could be the agent (he), the subordinates or someone who happens to witness the event described.

1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

B.2 Example Items

[Prompt 3]

Please provide the numerical rating along with the label defined for each question.

Practice question:

Consider the following utterance:

A: He would lure people into his car before driving off into the woods and stabbing them where no one could hear.

1. How would you rate the event described with regards to its connotation?
Event described: bring small gifts from trips abroad
Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning
1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)
2. How would a person affected by the event described feel?
Person affected: Could be the agent (he), the person being lured into the woods or someone who happens to witness the event described
1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

[Prompt 12]

Consider the following utterance:

- A: Why does everyone dance around acting like it's true?**
1. How would you rate the event described with regards to its connotation?
Event described: dance around acting like it's true
Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning
1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)
 2. How would a person affected by the event described feel?
Person affected: Could be the agent (everyone) or someone who happens to witness the event described
1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

B.3 Example Answers

[GPT-5]

EXP:

Consider the following utterance:

- A: Why does everyone go on acting like it's true?**
1. How would you rate the event described with regards to its connotation?
Event described: go on acting like it's true
Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning
1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)
 2. How would a person affected by the event described feel?
Person affected: Could be the agent (everyone) or someone who happens to witness the event described

1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

GPT-5:

1. 2 (somewhat negative)
2. 2 (somewhat unpleasant)

[Llama-3.3-70B]

EXP:

Consider the following utterance:

A: We went around talking about utopian visions.

1. How would you rate the event described with regards to its connotation?

Event described: go around talking about utopian visions

Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning

1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)

2. How would a person affected by the event described feel?

Person affected: Could be the agent (we) or someone who happens to witness the event described

1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

Llama-3.3-70B:

For the utterance:

1. 5 (very positive)
2. 5 (very pleasant)

[Gemini-2.5-Flash]

Chat logs were not saved due to an error.

[DeepSeek V3]

EXP:

Consider the following utterance:

A: We danced around humming.

1. How would you rate the event described with regards to its connotation?

Event described: dance around humming

Connotation: an associative, cultural and/or emotional meaning carried by a word or phrase, e.g. a feeling that is conveyed by a term which is not directly linked to its literal meaning

1 (very negative) – 2 (somewhat negative) – 3 (neutral) – 4 (somewhat positive) – 5 (very positive)

2. How would a person affected by the event described feel?

Person affected: Could be the agent (we) or someone who happens to witness

the event described

1 (very unpleasant) – 2 (somewhat unpleasant) – 3 (neutral) – 4 (somewhat pleasant) – 5 (very pleasant)

DeepSeek-V3:

3. How would you rate the agent’s intention behind the event described?

Agent’s intention: The perceived motive or purpose of the agent (we) in performing the action.

1 (very harmful) – 2 (somewhat harmful) – 3 (neutral) – 4 (somewhat beneficial) – 5 (very beneficial)

Your Answers:

1. 4 (somewhat positive)
2. 4 (somewhat pleasant)
3. 4 (somewhat beneficial)

APPENDIX C. DATA STRUCTURE AND STATISTICAL ANALYSIS

This appendix contains the variables found in the dataset, binary recoding rule, model formulae and packages used. To ensure replicability, the full dataset as well as the R workflow are available at the repository above (see Appendix A).

C.1 Variables

Table 3 Variables contained in the full dataset

Variable	Description	Values
Source	identifies human or LLM	Human, LLM
Group	identifies human or individual LLM	Human, Llama-3.3-70B, GPT-5, Gemini 2.5 Flash, DeepSeek V3
CASE	identifies individual human or LLM	16, 17, 25, 29, 30, Llama-3.3-70B, GPT-5, Gemini 2.5 Flash, DeepSeek V3
TypeNumber	identifies sentence type	S1, S2, S3, S4
Construction	identifies construction	goon, dancearound, goaround
Content	identifies construction + V2	goonacting, dancearoundhumming, goaroundtalking, dancearoundtelling, etc.
Question	identifies connotation/pleasantness questions	Connotation, Pleasantness
Rating	contains rating	1, 2, 3, 4, 5

C.2 Model Formulae

Table 4 Model formulae

Code	Description
<pre>dat_con <- dat_con %>% mutate(Rating_num = coerce_rating_numeric(Rating), con_neg12 = ifelse(Rating_num %in% 1:2, 1L, ifelse(Rating_num %in% 3:5, 0L, NA_integer_)))</pre>	Binary recoding rule (ratings 1-2 are recoded as 1, ratings 3-5 as 0); the same rules was applied to the pleasantness data
<pre>glm(formula = con_neg12 ~ Group + Construction, family = binomial("logit"), data = dat_con, method = "brglmFit")</pre>	Binary additive model (connotation data)
<pre>glm(formula = pls_neg12 ~ Group + Construction, family = binomial("logit"), data = dat_pls, method = "brglmFit")</pre>	Binary additive model (pleasantness data)
<pre>clm(Rating_ord ~ Group2 + Construction, data = dat_con_pool, link = "logit")</pre>	Ordinal regression model (cumulative link model with logit link) (connotation data)
<pre>clm(Rating_ord ~ Group2 + Construction, data = dat_pls_pool, link = "logit")</pre>	Ordinal regression model (cumulative link model with logit link) (pleasantness data)

ORCID

Victoria Krumpholz  <https://orcid.org/0009-0005-9639-3324>

REFERENCES

Bednarek, Monika. 2008. "Semantic preference and semantic prosody re-examined." *Corpus Linguistics and Linguistic Theory* 4(2): 119-139. <https://doi.org/10.1515/CLLT.2008.006>.

- Carchidi, Vincent J. 2024. “Creative minds like ours? Large Language Models and the creative aspect of language use.” *Biolinguistics* 18: e13507. <https://doi.org/10.5964/bioling.13507>.
- Christensen, Rune. 2023. “*ordinal*: Regression Models for Ordinal Data. R package version 2023.12-4.1.” Accessed November 1, 2025. <https://CRAN.R-project.org/package=ordinal>.
- Combs, Aidan, Diego Dametto, Christophe Blaison, et al. 2025. “Affective connotations according to LLMs: implications for meaning measurement and cultural bias.” *Cognition and Emotion*: 1–17. <https://doi.org/10.1080/02699931.2025.2568551>.
- Davies, Mark. 2008–. “The Corpus of Contemporary American English (COCA).” Accessed September 19, 2025. <https://www.english-corpora.org/coca/>.
- Dentella, Vittoria, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2024. “Testing AI on language comprehension tasks reveals insensitivity to underlying meaning.” *Scientific Reports* 14(1): 28083. <https://doi.org/10.1038/s41598-024-79531-8>.
- Goldberg, Adele E. 1995. *Constructions. A construction-grammar approach to argument structure*. University of Chicago Press.
- Hilpert, Martin. 2019. *Construction grammar and its application to English*. Edinburgh University Press.
- Hu, Krystal. 2023. “ChatGPT sets record for fastest-growing user base - analyst note.” *Reuters*, last modified February 2, 2023, <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.
- Jensen, Kim Ebensgaard. 2024. “Well, maybe you shouldn’t go around shaving poodles: collostructional semantic and discursive prosody in the go (a)round Ving and go (a)round and V constructions.” *Corpus Linguistics and Linguistic Theory* 21(3): 577-600. <https://doi.org/10.1515/clt-2024-0018>.
- Kosmidis, Ioannis. 2025. “*brglm2*: Bias Reduction in Generalized Linear Models. R package version 1.0.0.” Accessed November 1, 2025. <https://CRAN.R-project.org/package=brglm2>.
- Millière, Raphaël. 2024. “Language Models as Models of Language.” *arXiv*. <https://doi.org/10.48550/ARXIV.2408.07144>.
- Mortensen, David R., Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024. “Verbing Weirds Language (Models): Evaluation of English Zero-Derivation in Five LLMs.” *arXiv*. <https://doi.org/10.48550/arXiv.2403.17856>.
- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Sauerland, Uli, Celia Matthaei, and Felix Salfner. 2025. “Child vs. machine language learning: Can the logical structure of human language unleash LLMs?” *arXiv*. <https://doi.org/10.48550/arXiv.2502.17304>.
- Schütze, Carson T., and Jon Sprouse. 2014. “Chapter 3: Judgment Data.” In *Research Methods in Linguistics*, edited by Robert J. Podesva and Devyani Sharma, 27–50. Cambridge: Cambridge University Press.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4(43): 1686. <https://doi.org/10.21105/joss.01686>

Zhou, Shijia, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. “Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons.” *arXiv*. <https://doi.org/10.48550/arXiv.2403.17760>.