

Using LLMs for experimental stimulus pretests in linguistics. Evidence from semantic associations between words and social gender

Christian Lang and Franziska Kretzschmar and Sandra Hansen

Department of Grammar, Leibniz Institute for the German Language, Mannheim, Germany
lang@ids-mannheim.de, kretzschmar@ids-mannheim.de, hansen@ids-mannheim.de

Abstract

Whether large language models (LLMs) can validly complement or substitute human participants in experimental research remains an open question. Focusing on language cognition, we assess the suitability of GPT-4o and LLaMA 3.1 models (70B Instruct and 8B Instruct) for performing a semantic-association task in German. LLMs labeled noun phrases by social-gender association and rated association strength, mirroring a human participant task. Overall, LLM ratings aligned with human data, but item-level analyses revealed systematic deviations in response patterns.

1 Introduction

The increasing abilities of large language models (LLMs) to generate human-like output has initiated empirical investigations on whether LLMs may be a valid experimental data source either in addition to humans or even replacing them (e.g., Aher et al., 2023; Argyle et al., 2023; Cai et al., 2024; Kuribayashi et al., 2024; Misra et al., 2020; Zhu et al., 2023). The main arguments for using LLM-generated data over human data were that LLMs do not suffer from humans' adaptations to an experimental procedure (e.g., fatigue or learning effects, cf. Digutsch and Kosinski, 2023), and, as a research-supporting tool, LLMs would mitigate monetary and time costs at preparatory stages in one's research project such as in pilot studies or item pre-tests (cf. Dillion et al., 2023; Trott, 2024).

Strikingly, the results of the existing comparisons of authentic natural and artificial datasets across varying tasks suggest that despite the qualitative similarity of LLM-based and human-based results overall, LLM-generated data often show a tendency to over- or underestimate aggregate measures of human data. As argued (and empirically tested) in detail by Trott (2024), this may lead to differences in magnitude between LLM-based and

human-based results. Further research is needed to analyze the robustness of this pattern, i.e. to what extent LLM-based and human-based data diverge or converge at the (aggregated) condition and item level depending on LLMs, tasks and languages.

Here, we compare data from three LLMs and a human sample presented with the same task. Our use case is an item pre-test, a necessary step in experimental research to control stimuli for implicit biases that may have undesirable effects on the results of the main study. In its simplest form, a separate participant sample would perform some task with a set of items. Based on the aggregated pre-test results, the researcher would then select items for the main study. Our research question is to what extent different LLMs are suitable for this type of item pre-test. We analyze LLMs performance vis-à-vis human behavior in a semantic-association task involving social-gender associations in German.

We investigate the stereotypical association of the lexical meaning of a noun phrase (NP) with a person of female or male gender (i.e., social gender, see Lewis and Lupyan (2020) for an overview of its cross-linguistic influence, and Esaulova et al. (2017) for the impact of social-gender associations on sentence comprehension in German).¹ If included in a pre-test, participants would typically have to label (i.e., associate) words for social gender and/or rate the association strength – which we implemented as a task in our study.

We consider this task an ideal testing ground to compare results from LLMs with human participants, because LLMs basically act as models of distributional semantics (Enyan et al., 2024). This makes them predestined for questions of (semantic) association, or context-sensitive phenomena more

¹A prominent example is the NP *die Flugbegleitung* [flight attendant] that is typically associated with a female person in German (Esaulova et al., 2017). However, other role descriptions for people may have associations that are less obvious to language users, e.g. *genius*, *authority* or *movie star*.

generally. The extensive training data of LLMs contain data points relevant for our (or any) task. When prompted neutrally – i.e., not limited to a specific persona –, the output of the LLMs provides an aggregation of these relevant data points that is representative of the training data. We test to what extent this aggregation is comparable to human data at the condition and item level.

In general, our study adds new data from an hitherto unstudied experimental task to the literature, as well as data from a target language other than English. Moreover, unlike the pertinent prior studies with large lexical databases and mainly ChatGPT models (see Section 2), our work examines 1) a novel semantic dimension—social-gender associations, 2) a more ecologically valid item set regarding size and composition, and 3) models from different families.

2 Related Research

For English psycholinguistic databases, it has been found that LLMs and humans generate similar data when condition-based aggregations are compared (e.g., Digutsch and Kosinski, 2023; Trott, 2024). However, it has also been found that effect sizes vary, i.e. although LLMs show relative differences between conditions that are similar to humans, absolute values for the measures per condition differ (even strikingly so for some tasks, cf. Cai et al., 2024; Trott, 2024). For example, in their study on semantic-association priming, the difference in effect size led Digutsch and Kosinski (2023) to conclude that priming due to semantic similarity (overlap in number of semantic features) has a stronger impact on LLM performance than semantic association (co-occurrence of words). Similarly, Trott (2024) found in a series of regression analyses that many predictor variables show changes in effect size depending on the type of data generation, with some of them suggesting that information on semantic categories are more important to humans than to LLMs. Thus, quantitative differences across conditions suggest qualitative differences in how humans and LLMs use semantic information.

3 Data base and methodology

We provide supplementary materials including the item list, full LLM prompts, a sample workflow to generate and process the LLM data, aggregated data, and the analysis script here: https://osf.io/ezy9p/?view_

[only=c72db52d721c49fb9be492bd4b914d10](https://osf.io/ezy9p/?view_only=c72db52d721c49fb9be492bd4b914d10).

3.1 Dataset

From the literature on grammatical gender marking and social gender in German (Thurmair, 2006; Klein, 2022; Kopf, 2022; Schoenthal, 1989), we identified NPs from each grammatical-gender category (masculine, feminine, neuter) for which social gender may vary between male, female or no clear association.² From the resulting list, we excluded pejoratives, animal names used as rare metaphors for humans, words for which there was a highly salient suffixed word form to indicate a female person (such as *Passagierin* [female passenger]), ambiguous and infrequent words. If a base word allowed several compounds, we did not use more than four (e.g., *star*: *movie star*, *pop star*) to match the number of experimental lists (see Section 3.2.2). To balance the distribution of grammatical gender, we added 19 NPs not covered in the cited literature. The final list included 116 NPs: 30 feminine, 47 masculine, 36 neuter, and 3 with varying grammatical gender (e.g., *der/das Balg* [annoying child]).

3.2 General experimental task

The semantic-association task consisted of two steps, which were identical for LLM prompts and task instructions for humans. First, LLMs and humans had to label NPs for their associated social gender, choosing either the label *Person männlichen Geschlechts* [person of male gender], *Person weiblichen Geschlechts* [person of female gender], or *keins von beiden* [neither]. Second, if an item was labeled as person of male or female gender, we asked the LLMs/human annotators to rate the association strength on a scale from 1 (very low) to 5 (very high). The following subsections describe the specifics of data collection for LLMs and the human sample, respectively.

3.2.1 LLM-Generated Assessments

We applied three different LLMs —GPT-4o (a widely adopted proprietary model) and LLaMA 3.1 (70B and 8B, open-source models differing in size)— to label 116 NPs for social gender using zero-shot prompting (see supplementary materials).

We accessed LLaMA models via ChatAI (Doosthosseini et al., 2024) and GPT-4o through OpenAI’s commercial service, using OpenAI’s chat

²For example, the grammatical gender of *die Aushilfe* [temporary staff] is feminine, but the associated social gender can be male, female or neither. Note that the NP selection was part in stimulus creation for a future reading study.

completion API for all models. Each labeling was initiated through a separate API call to prevent contextual carryover. The system prompt (developer role, in English) defined a neutral assistant persona and structured output for easier processing, aligning with the pre-test goal of a general, condition-level assessment. The task prompt (user role) was delivered in German and templated to ensure consistent wording across all 116 noun phrases. We chose this language combination to maximize clarity (via English system prompt) and linguistic relevance (via German task prompt).³

Due to the non-deterministic nature of LLMs, outputs generated with the same prompt can differ from one another.⁴ Hence, and in line with previous research (Elman, 1990; Wilcox et al., 2024), we used each LLM to generate multiple labels and ratings (N = 33) per item, corresponding to the number of participant labels and ratings per NP (see Section 3.2.2). We kept the temperature at 0 to make LLM outputs as deterministic—and thus as reproducible—as possible.

3.2.2 Study with human participants

The experiment was designed to be as similar as possible to the LLM-generated assessments, with two modifications following standards in psycholinguistic experiments. First, items were allocated to 4 lists of 29 NPs each to rule out fatigue or learning effects of human participants; each participant was randomly assigned to one of the lists. The lists were compiled according to the following criteria: a) nearly even distribution of grammatical gender across lists, and b) no more than two items with the same word stem (e.g., *star* and *movie star* etc.) per list. Moreover, to prevent a list from containing a majority of items with a clear bias towards one social gender, we used the LLM ratings to allocate NPs to lists with a near-equal distribution regarding LLM-based gender labels. Second, items were presented with a unique randomized order per participant to prevent sequence effects.

The study was run on the survey platform *Unipark* (<https://www.unipark.com>). 132 participants (52 females, 76 males, 3 diverse, 1 not specified; age range: 22-75) were recruited with *Prolific* (<https://www.prolific.com>) and received mon-

³English prompts often yield better performance (cf. Schulhoff et al., 2025, p. 20), while the task required attention to German-specific features such as grammatical gender.

⁴Even with temperature 0 and a fixed seed, the OpenAI API produces varying outputs; an initial 10-run test showed noticeable variation in NP labels.

Annotator	m	n	f	unclear
Human annotators	39	65	11	1
GPT-4o	31	69	16	0
LLaMA 3.1 70B	36	67	13	0
LLaMA 3.1 8B	111	0	5	0

Table 1: Distribution of majority vote labels by annotator (LLM or human annotators)

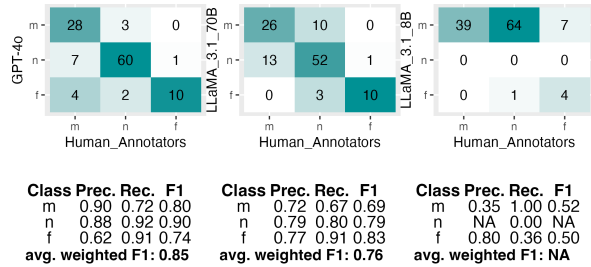


Figure 1: Top: Confusion matrices with overlap of associative labeling per category, based on majority votes; N = 115 (we excluded the NP with unclear majority vote). Y-axis: LLM labels, X-axis: human labels (reference). Bottom: macro average of precision, recall and F1

etary compensation for their participation. There were no missing or erroneous responses, resulting in 33 data points per NP/item.

4 Results

In the following, we use the human annotators as a benchmark against which the LLMs are compared. Table 1⁵ shows the majority-vote distribution (most frequent label across 33 trials) for humans and LLMs. All NPs had a majority vote, except *Scheusal* [beast] with human annotators, which showed a 14–14 tie between *person of male gender* and *neither*.

The distributions are very similar, with the label *neither* being the most frequent and the label *person of female gender* being the least frequent. There appear to be only small differences between LLM-generated and human labels (e.g., the more frequent assignment of the label *person of female gender* by GPT-4o). An exception is LLaMA 3.1 8B, which deviates fundamentally in that it almost exclusively assigns the label *person of male gender* to the NPs and never assigns the label *neither*.

The confusion matrix in Figure 1 shows the alignment between the LLM labels and human-based labels, along with performance metrics per

⁵Here and below: *m*: *person of (social) male gender*; *n*: *neither*; *f*: *person of (social) female gender*.

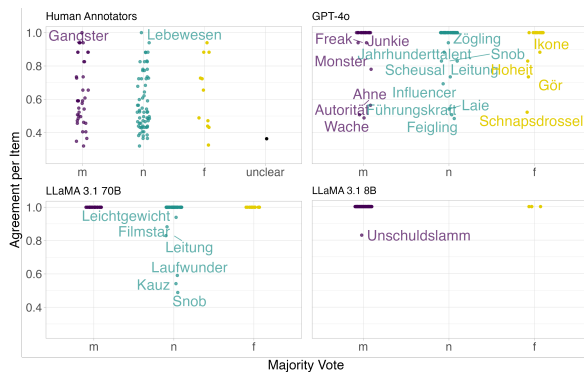


Figure 2: Agreement per item between trials/iterations (N = 116 items); NP names shown: humans agreement = 1; LLMs agreement < 1

gender category (macro average of precision, recall, and F1 score). Again, GPT-4o shows the most human-like performance, whereas LLaMA 3.1 8B deviates most from human annotations. Although Table 1 and Figure 1 suggest reasonable alignment between GPT-4o (and to a lesser extent LLaMA 3.1 70B) and human annotators, further analysis reveals a more nuanced picture.

We calculated the observed percent agreement P_i for each item i using Equation 1⁶ (cf. Fleiss, 1971, p. 379):

$$P_i = \sum_{j=1}^k \frac{n_j(n_j - 1)}{n(n - 1)} \quad (1)$$

Figure 2⁷ shows the percent observed agreement for every item for humans and LLMs. Strikingly, LLMs produce perfect agreement (agreement = 1) across all 33 iterations for the majority of items, whereas only two items (*Gangster* = ‘person of male gender’ and *Lebewesen* [living being] = ‘neither’) show perfect agreement among humans. Within the group of LLMs, GPT-4o exhibits a greater number of items with agreement < 1 than the other models, and LLaMA 3.1 8B shows disagreement for only one item (*Unschuldslamm* [innocent lamb]).⁸

Figure 3 illustrates the distribution of LLM-generated labels for the items with an agreement

⁶ k represents the total number of labels (male gender, female gender, neither: $k = 3$), n is the total number of annotators per item ($n = 33$), and n_j denotes the number of annotators that assigned the item to label j .

⁷All images are available in higher resolution in the supplementary materials.

⁸An interactive visualization showing label distributions per item and the proximity between LLMs and human annotators is available at https://wdsmuek.shinyapps.io/vote_viewer/.

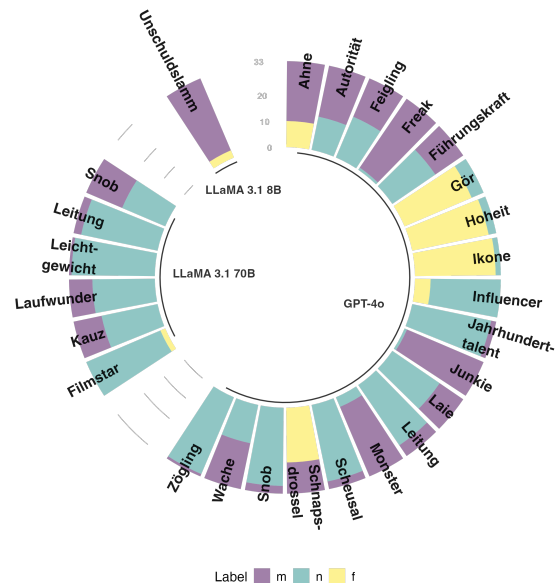


Figure 3: Distribution of labels for items with LLM agreement < 1

< 1. It is noticeable that the labels *person of male gender* and *person of female gender* were only assigned to the same item in three cases (*Ahne* [ancestor], *Schnapsdrossel* [whino], *Unschuldslamm* [innocent lamb]). In all other cases, the agreement value < 1 results from the combination of either *person of male gender* or *person of female gender* with *neither*. This does not apply to the human annotators who assigned all three labels to several items, as is visible in Figure 4.

Figure 5 shows the distribution of association-strength values 1 (very low) to 5 (very high). Both humans and LLMs rated their labels with high association strength, with a mean value of 4; this pattern was more pronounced for the male gender category. However, only humans make use of the entire scale, LLMs rarely label with (very) low association strength, as reflected in smaller standard deviation (SD) values: humans ± 1.17 , GPT-4o ± 0.5 ; LLaMA 70B ± 0.78 , LLaMA 8B ± 0.18 .⁹

5 Discussion

We compared LLMs output with human behavior in a semantic-association task, where German NPs had to be labeled for their association (strength) with a social gender. The key findings are: 1) There is substantial variation between LLMs, with

⁹We further analyzed the summed log-probabilities of label sub-tokens in relation to agreement and association strength. GPT-4o showed a positive correlation with agreement, and both GPT-4o and LLaMA 3.1 70B with association strength. See supplementary materials for visualizations.

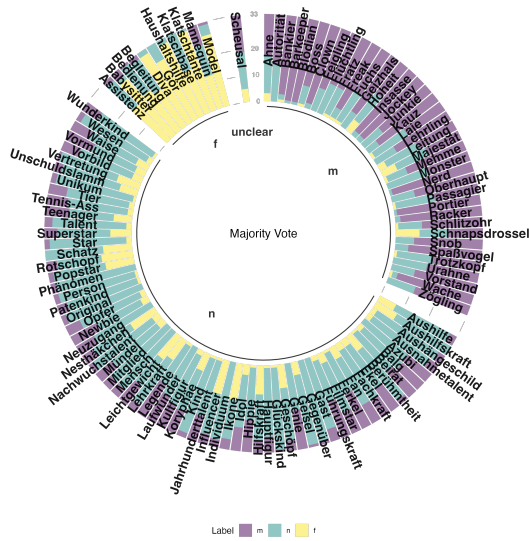


Figure 4: Distribution of labels for items with human annotator agreement < 1

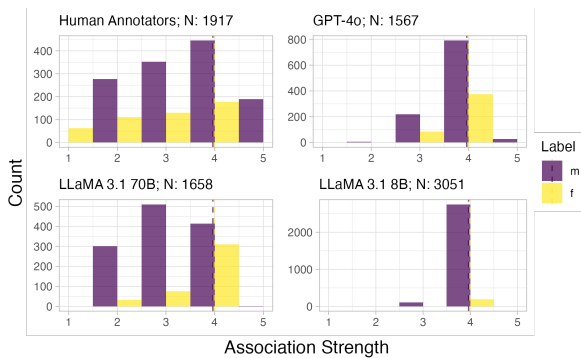


Figure 5: Distribution of association strength. Dashed lines represent the mean per gender category.

the smallest model (LLaMA 3.1 8B) showing the least human-like output. 2) In terms of social-gender labels, the two larger models produce aggregate output (majority votes) comparable to humans, hence there is aggregation/condition-level similarity (Fig. 1). 3) At the item level, the distribution of raw data points suggests differences between trials in that the human data exhibit larger variation per item compared with the non-human-like consistency of LLMs (i.e., perfect agreement, Fig. 2). This aligns with differences in scale effects, as LLMs, in contrast to humans, label NPs only with higher, but not with lower association strength (Fig. 5) – which may be plausible given LLM-internal probabilities (see footnote 9).

Our results support previous findings that associative semantic relations may be treated differently by LLMs and humans (Digutsch and Kosin-

ski, 2023), suggesting that LLMs task performance is qualitatively different from human responses at the item level. While, across trials for NPs with an agreement < 1 , human responses may include any of the three social-gender labels for an NP, LLMs use only two labels (Figs. 3 & 4). One reason may be that humans use context information based on (inter alia) differences in exposure, salience or individual significance, thereby promoting inter- and intraindividual variation. Hence, different people may have different associations for a *pop star*, for example. When prompted with the same neutral prompt as in our study, these factors cannot exert any influence for LLMs context use. Thus, neutrally prompted LLMs may capture aggregation-based tendencies of social-gender associations across varied contexts, but capturing individual variation requires at least carefully designed personas, if not other ways of data generation (see Franke et al., 2024; Murthy et al., 2025).

6 Summary

We find similar condition-level (aggregated) results for LLMs and humans, while the underlying distribution of responses differs in a semantic-association task. This supports conclusions that LLMs may not, in every task, serve as a substitute for human data beyond qualitative (condition-based) effect reports (Franke et al., 2024; Murthy et al., 2025; Trott, 2024). For our use case, experimental pre-tests, LLMs may be useful for condition-based assessments regarding item selection. Yet, when pre-test data are used with other statistical methods (e.g., power analysis, simulation), the current approach to LLMs use may not add insightful data. In such cases, substituting human data with LLM data may cause Type-M errors, affecting replicability across AI and human datasets (for discussion of Type-M error in (psycho)linguistics, cf. Vasishth, 2023).

Limitations

Our study relies on prompt-based methods and is thus subject to the limitations identified by Hu and Levy (2023). We employ a single prompt, despite evidence that minor prompt variations can significantly affect outputs (cf. Schulhoff et al., 2025). It is also an open question how individual variation can be accurately captured. Finally, the use of only three language models limits the generalizability of our findings.

References

- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. [Do large language models resemble humans in language use?](#) In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.
- Jan Digitsch and Michal Kosinski. 2023. [Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans](#). *Scientific Reports*, 13(1):5035.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. [Can AI language models replace human participants?](#) *Trends in Cognitive Sciences*, 27(7):597–600. Publisher: Elsevier.
- Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian M. Kunkel. 2024. [Chat ai: A seamless slurm-native solution for hpc-based services](#). *Preprint*, arXiv:2407.00110.
- Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.
- Zhang Enyan, Zewei Wang, Michael A. Lepori, Ellie Pavlick, and Helena Aparicio. 2024. [Are LLMs Models of Distributional Semantics? A case study on quantifiers](#). *Preprint*, arXiv:2410.13984.
- Yulia Esaulova, Chiara Reali, and Lisa von Stockhausen. 2017. Prominence of Gender Cues in the assignment of Thematic Roles in German. *Applied Psycholinguistics*, 38(5):1132–1172.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Michael Franke, Polina Tsvilodub, and Fausto Carcassi. 2024. [Bayesian Statistical Modeling with predictors from LLMs](#). *Preprint*, arXiv:2406.09012.
- Jennifer Hu and Roger Levy. 2023. [Prompting is not a substitute for probability measurements in large language models](#). *Preprint*, arXiv:2305.13264.
- Andreas Klein. 2022. [Wohin mit Epikoina? – Überlegungen zur Grammatik und Pragmatik geschlechtsindefiniter Personenbezeichnungen](#). In Dabriele Diewald and Damaris Nübling, editors, *Genus – Sexus – Gender*, pages 135–189. de Gruyter, Berlin.
- Kristin Kopf. 2022. [Ist Sharon Manager? Anglizismen und das generische Maskulinum](#). In Dabriele Diewald and Damaris Nübling, editors, *Genus – Sexus – Gender*, pages 65–103. de Gruyter, Berlin.
- Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. [Psychometric predictive power of large language models](#). *Preprint*, arXiv:2311.07484.
- Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4:1021–1028.
- Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. [Exploring BERT’s sensitivity to lexical cues using tests from semantic priming](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.
- Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. [One fish, two fish, but not the whole sea: Alignment reduces language models’ conceptual diversity](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gisela Schoenthal. 1989. Personenbezeichnungen im Deutschen als Gegenstand feministischer Sprachkritik. *Zeitschrift für Germanistische Linguistik*, 17(3):296–314.
- Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yin-heng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncarencu, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. [The prompt report: A systematic survey of prompt engineering techniques](#). *Preprint*, arXiv:2406.06608.
- Maria Thurmair. 2006. Das Model und ihr Prinz. Kongruenz und Texteinbettung bei Genus-Sexus-Divergenz. *Deutsche Sprache*, 34:191–220.
- Sean Trott. 2024. Can Large Language Models help augment English psycholinguistic datasets? *Behavior Research Methods*, 56:6082–6100.
- Shravan Vasishth. 2023. Some right ways to analyze (psycho)linguistic data. *Annual Review of Linguistics*, 9:273–291.
- Ethan Gottlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. [Can ChatGPT reproduce human-generated labels? A study of social computing tasks.](#) *Preprint*, arXiv:2304.10145.