

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

LLMs Struggle With Negation, but so do Humans - a Two-step Simulation Approach

Permalink

<https://escholarship.org/uc/item/92v7f02h>

Journal

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

Authors

Döcker, Lenny Tim

Erens, Kevin Nick

Kaup, Barbara

et al.

Publication Date

2026

Copyright Information

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

Peer reviewed

LLMs Struggle With Negation, but so do Humans - a Two-step Simulation Approach

Lenny Döcker (lenny.doecker@student.uni-tuebingen.de), Kevin Erens, Barbara Kaup & Francesca Capuano

Department of Psychology, University of Tübingen

Abstract

While substantial research has addressed the challenges of negation processing in humans, the parallels between human cognitive models and large language models (LLMs) in this regard remain less explored. This paper investigates how GPT-4o processes negation, drawing comparisons to human negation processing, particularly the two-step model of negation. In two experiments using image-sentence pairs, we assess the model's ability to handle affirmative and negated sentences. Our findings reveal that, like humans, GPT-4o struggles more with negation, exhibiting higher rates of incorrect inferences after negation, and a systematic bias toward completions associated with the negated state of affairs. This study highlights the qualitative similarities between AI and human processing of negation, offering insights into the limitations of current LLMs and suggesting future directions for improving their cognitive alignment with more complex linguistic constructs.

Keywords: Negation; Multimodal Models; Language Processing; Two-step Model

Introduction

Surprisingly despite the continuous advances in NLP, models like BERT or the OpenAI GPT series have been singled out for their deficiencies in processing negation (e.g., Ettinger, 2020; Kassner and Schütze, 2019). Solutions to address this issue often require explicit engineering rather than developing naturally as model performance improves: while general improvements in architecture and training data have led to progress in many language tasks, negation remains an area where spontaneous advancements are limited (e.g., García-Ferrero et al., 2023; Jang et al., 2023; Khandelwal & Sawant, 2019). This persistent challenge highlights the need to examine how models handle complex linguistic constructs, such as negation - not merely from a performance standpoint but also in terms of cognitive alignment and interpretive depth - to gain insights into both artificial intelligence and human language comprehension. However, research in NLP has often prioritized quantitative performance metrics, perhaps neglecting the more qualitative aspects that reflect the long-standing challenges humans face with negation processing.

Negation processing has long been associated with increased cognitive costs in the psycholinguistics literature, a phenomenon documented as the "negation effect" (Carpenter and Just, 1975; Clark and Chase, 1972; Wason, 1959; for an overview, see Kaup and Dudschig, 2020). One explanation for this effect is rooted in experiential views of language

processing (Johnson-Laird, 1983; Zwaan and Madden, 2005), which propose that we mentally simulate linguistic input as if we were directly experiencing it. However, when something is negated (e.g., "There is no eagle in the sky"), we can't "experience" the absence of something in the same way as we experience its presence. This raises the question: how can our minds represent something that is not there?

One accredited explanation is proposed by the *two-step model* of negation processing. According to this model, upon encountering a negated sentence, individuals first construct a mental image of the negated state of affairs. For example, hearing "There was no eagle in the sky", the brain first constructs a mental image of an eagle in the sky, as though the sentence were affirming the eagle's presence. In the second step, the brain updates this image to reflect the actual state of affairs - that is, the sky without an eagle. Evidence for the first step comes from studies such as Kaup et al. (2007), where participants were presented with negated sentences such as "There was no eagle in the sky" and then shown images shortly afterward (250 ms), depicting either an eagle with wings spread or with wings folded (see Figures 1a and 1b). Participants were asked to identify whether the eagle had been mentioned in the preceding sentence. They reacted faster to the image of the eagle with wings spread out, indicating that, initially, they were simulating the negated state of affairs. Kaup et al. (2006) tested the second step by examining response patterns at later latencies (750 ms and 1500 ms). Using contradictory predicates like "The door was open" vs. "The door was not open" and "The door was closed" vs. "The door was not closed", participants named objects in pictures that either matched or mismatched the described state. At the 750 ms delay, a match effect with the actual state of affairs was observed for affirmative sentences, but not for negative ones. However, at the 1500 ms delay, the match effect emerged for negative sentences, suggesting that the participants' mental representations had by then been updated to reflect the negated state, while the match effect for affirmative sentences had disappeared. Together, these results suggest that understanding negation is cognitively taxing because the brain initially processes the negated sentence as if it were affirmative, simulating what the sentence negates. Only after this initial step does the brain move on to process the actual state.

Although extensive research has highlighted the challenges humans face with negation, the specific mechanisms through which language models struggle in this domain remain less



Figure 1: Example images matching (a) or mismatching (b) the simulation of the negated state of affairs in "There is no eagle in the sky".

understood. This paper seeks to explore how GPT-4o, a multimodal model that integrates both text and image processing, processes negation in the context of visual input, investigating whether the model’s responses exhibit patterns consistent with those predicted by the two-step model of negation. To address this, we present images paired with sentence beginnings that must be completed to align with the depicted visuals, mirroring the experimental approach of Kaup et al. (2007). Affirmative and negated sentences are evaluated for their alignment with the images, and responses are categorized as correct, incorrect, or neutral. In our experiments, we aim to evaluate two hypotheses: first, that negated sentences will yield more false completions than affirmative ones, indicating a negation effect; second, that we would observe a predominance of false inferences based on the negated states rather than the actual circumstances, suggesting GPT-4o’s inability to adequately represent the second step of negation processing.

Through this qualitative exploration, we aim to gain insights on the mechanisms behind GPT-4o’s challenges with negation and enhance our understanding of its input image capabilities. Our study investigates not only whether the model encounters difficulties with negation but also whether it parallels human struggles with negation at a more qualitative level. By examining the model’s responses in relation to human processing, we can gain insights into the similarities and differences in understanding negation, fostering a deeper comprehension of how both systems approach the complexities of negation.

Experiment 1

Experiment 1 aims to examine whether GPT-4o exhibits a "negation effect" similar to humans when processing negated statements. The model was presented with images of common objects and given sentence fragments containing either an affirmative or negated statement, followed by two possible completions. By analyzing the model’s choices, we aimed to assess its ability to correctly integrate negation with visual context. The key hypothesis is that GPT-4o would demonstrate a higher rate of incorrect completions when processing negated sentences, suggesting a cognitive-like difficulty in suppressing the initial affirmative interpretation before updating its understanding to reflect the negated state.

Materials

We used 56 images of objects against a white background. We based these images on the ones used in the study by Schütt et al. (2023). Two of them were, for example, an egg in its shell and a fried egg. We created affirmative and negated sentence beginnings with two word options as a completion (e.g., "There was no/an egg in the (refrigerator/skillet)."). The sentences follow the structure found in Kaup et al. (2007). The images fall under copyright restrictions, but the sentential items are openly available on OSF.

Procedure

We conducted the experiments with ChatGPT-4o and used 10 chat windows. At the start of each chat we gave the instruction to complete sentence beginnings referring to the accompanying image by selecting one of the provided word options.

Each chat contained every image paired with both an affirmative and a negated sentence beginning, following a Within-Subject-Design (for an example, see Table 1). The 112 image-sentence pairs were presented in random order. The order of the word options in brackets was also randomized.

The response was then categorized as aligning with the image (correct), or not aligning with the image (incorrect). For affirmative sentences like "There was an egg in the (refrigerator/skillet).", paired with an image of a fried egg, "skillet" would be a correct and "refrigerator" an incorrect completion. For negated sentences like "There was no egg in the (refrigerator/skillet).", paired with the same image, "skillet" would be an incorrect and "refrigerator" a correct completion. We calculated the proportion of these ratings for each test run and polarity level (affirmative/negated).

It should be noted that our *correctness* criterion is based on pragmatic assumptions rather than strict logical inconsistency. In fact, some image-sentence pairings that we label as incorrect might actually be semantically or logically acceptable under different interpretations (e.g., a fried egg could be in a refrigerator, and an uncooked egg could be placed in a skillet). The approach follows the methodology in Kaup et al. (2007), which also assume a binary alignment between linguistic input and visual stimuli. For example, just as Kaup et al. treated an image of an eagle with folded wings as inconsistent with the sentence "There is an eagle in the sky", our criteria required that an image match the most salient interpretation of the linguistic description. However, it is possible that our items introduce greater interpretative flexibility than those used by Kaup et al. (2007), raising the question of whether they would replicate their findings in the first place. This issue will be addressed in Experiment 3.

Results

The results are presented in Figure 2. A two-sided paired t-test showed that the probability for an incorrect response was higher for negated sentences ($M = 20.42$, $SD = 5.72$) than for affirmative sentences ($M = 9.17$, $SD = 3.02$), $t(9) = 5.80$, $p < .001$, 95% CI [6.86, 15.64].

Table 1: Example pictures-sentences pairs. The fried egg (top left) refers to the state of affairs of an egg in a skillet. The egg in its shell (bottom left) refers to the state of affairs of an egg in a refrigerator.

Affirmative	Negative
 Exp 1: There was an egg in the (refrigerator/skillet). Exp 2: There was an egg in the ____	 Exp 1: There was no egg in the (refrigerator/skillet). Exp 2: There was no egg in the ____
 Exp 1: There was an egg in the (refrigerator/skillet). Exp 2: There was an egg in the ____	 Exp 1: There was no egg in the (refrigerator/skillet). Exp 2: There was no egg in the ____

This pattern supports the presence of a negation effect, with the model showing a marked decrease in performance when processing negated compared to affirmative descriptions.

Experiment 2

Experiment 2 extends the results of Experiment 1 by removing the predefined word choices for sentence completion, allowing GPT-4o to generate responses more freely. This modification provides insight into whether the model defaults to a particular type of inference when handling negation, as well as whether it could independently integrate visual context to guide its completions. By introducing a neutral category for ambiguous responses, the study explores whether errors primarily stemmed from the model failing to process negation properly or from general uncertainty in sentence completion.

Materials

The same images as in Experiment 1 were used. In contrast to Experiment 1, we created affirmative and negated sentence beginnings without continuation options (e.g., "There was an/no egg in the").

Procedure

The procedure of Experiment 2 was identical to Experiment 1, with the exception that ChatGPT-4o was instructed to complete sentence beginnings referring to the accompanying image by selecting one word freely. Example image-sentence pairs are shown in Table 1.

Responses that could not be assigned to the correct or incorrect categories were rated as *neutral*. For affirmative/negated sentences like "There was an/no egg in the", paired with an

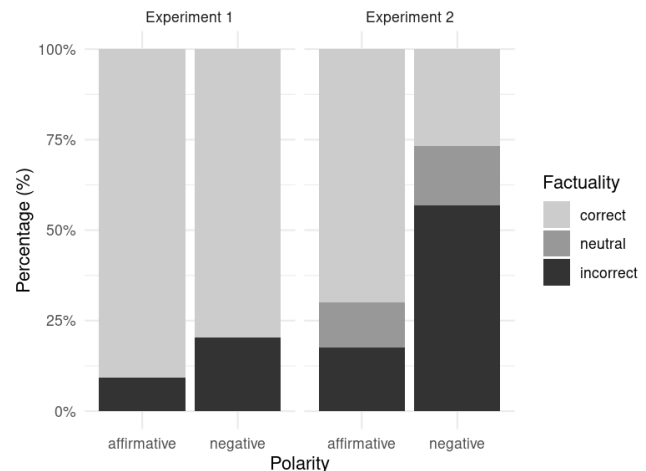


Figure 2: **Experiments 1 and 2** - Percentage of incorrect, correct and neutral responses.

image of a fried egg, "kitchen" or "house" would be neutral completions because the answer does not provide any information about whether the egg is in its shell or fried.

We calculated the proportion of these ratings for each test run and polarity level (affirmative/negated).

Results

The results are presented in Figure 2. A two-sided paired t-test showed that the probability for an incorrect response was higher for negated sentences ($M = 56.90$, $SD = 5.85$) than for affirmative sentences ($M = 17.51$, $SD = 4.77$), $t(9) = 19.69$, $p < .001$, 95% CI [34.86, 43.92].

Moreover, for negated sentences, incorrect responses occurred above chance level, $t(9) = 3.73$, $p = .005$, 95% CI [52.71, 61.08]. Thus, when GPT-4o was allowed to generate completions freely, negation did not merely lead to greater variability or uncertainty. Instead, the model more often produced completions that were incompatible with the actual visual context.

To further examine whether the effect reflected systematic false inferences rather than underspecified completions, we compared the polarity-related increase in incorrect responses with the corresponding increase in neutral responses. The increase in incorrect responses from affirmative to negated sentences ($M = 39.39$, $SD = 6.33$) was significantly larger than the increase in neutral responses ($M = 3.86$, $SD = 4.59$), $t(9) = 12.56$, $p < .001$, 95% CI [29.14, 41.93]. This suggests that the model's difficulty with negation was not primarily driven by a failure to choose a specific completion. Rather, negation systematically shifted responses toward completions associated with the negated state of affairs.

Experiment 3

In response to concerns about the direct comparability of our findings to human performance, we additionally conducted a replication of Kaup et al. (2007)'s experiment, using the same

image-sentence pairs administered to the model in Experiment 1. By explicitly incorporating human data, we eliminate the concern that our comparison relies on different experimental items from previous studies.

The study was preregistered.

Participants

A total of 80 native English speakers (45 females, 34 males and 1 other; age: $M = 40.95$, $SD = 14.24$) participated in the experiment and received £2.50 as compensation (in compliance with the minimum wage guidelines of the Psychology Department at the University of Tübingen). The data was collected via Prolific (www.prolific.com). Participants conducted the experiment using their personal computers. Participants provided informed consent.

Materials

The same images as in Experiments 1 and 2 were used for the experimental trials. Unlike in Experiments 1 and 2, the 56 experimental sentences were all negated and complete sentences (e.g., "There was no egg in the skillet." and "There was no egg in the refrigerator."). With respect to the corresponding images - egg in its shell and fried - this resulted in four possible image-sentence pairs for one experimental item (2 sentences $\times$ 2 images).

A total of 56 filler images were used paired with 42 affirmative and 14 negated filler sentences. Only 14 of the affirmative sentences were followed by objects mentioned in the preceding sentence. The rest of the filler sentences were accompanied by new images picturing objects which were not mentioned in any sentence. Following the design of Kaup et al. (2007), comprehension questions were included for 24 of the filler sentences. Furthermore, a set of 12 additional image-sentence pairs was employed for practice trials, accompanied by 4 comprehension questions. Half of the practice sentences were negated, while the other half were affirmative.

Procedure

The procedure followed Kaup et al. (2007)'s Experiment 1. Participants were instructed to read a sentence and then decide whether the following image showed an object mentioned in the preceding sentence. They were told to respond as quickly and accurately as possible using the j-key to give a "yes" and the f-key to give a "no" response. The instruction was followed by the practice trials.

The participants were evenly assigned to four lists, counterbalancing items and conditions. Every list included a different one of the four possible versions of the experimental items. For each list this resulted in a total of 28 negative experimental sentences that were paired with images that required a "yes" response. In addition, each participant saw 14 negative filler sentences that were paired with images requiring a "no" response, 28 affirmative filler sentences that were paired with images requiring a "no" response, and 14 affirmative

filler sentences that were paired with images requiring a "yes" response.

Each trial started with the sentence. Participants chose when to continue by pressing the spacebar. A fixation point appeared for 250 ms before the image was shown. Below the image participants were given the two response options: A green button labeled "J (yes)" and a red button labeled "F (no)". No feedback was provided on the accuracy of the response. Some trials were followed by comprehension questions that could be answered in the same way. Incorrect answers triggered "WRONG" as feedback. It took about 15 minutes to complete the experiment.

Results

The analysis followed the approach used in Kaup et al. (2007) to ensure comparability. Responses exceeding 3,000 ms or shorter than 300 ms, as well as those falling outside two standard deviations from the participant's mean within each condition, were excluded.

Reaction times (RTs) for experimental trials were analyzed using a 2 (depicted situation: negated vs. other) $\times$ 4 (list) repeated measures analysis of variance (ANOVA) for both the by-participant and by-item analyses. Only correct responses were included. In line with best practices, we also pre-registered a more modern approach using linear mixed-effects models, which are now the standard in psycholinguistic research.

The results of the ANOVA showed significantly faster responses when the depicted situation matched the negated situation compared to when it did not, $F_1(1, 74) = 7.44$, $p < .01$ (see Figure 3). The effect did not reach significance in the by-picture analysis, $F_2(1, 50) = 2.54$, $p = .12$. However, the LMM analysis, which accounts for both participant- and picture-level variance, confirms a significant effect of depicted situation on RTs ($\beta = 29.27$, $SE = 11.53$, $p < .05$). Response accuracy did not show a significant effect of depicted situation, either in the ANOVA or in the LMM analysis. However, errors were extremely rare, with most participants ($N = 59$) achieving perfect accuracy. Given this ceiling effect, the accuracy analysis is not particularly informative.

These findings support the hypothesis that comprehenders initially simulate the negated situation when processing a negative sentence. Specifically, they appear to encode the object's shape in the negated context, leading to faster responses when the picture matches this shape compared to when there is a mismatch.

Most importantly, we replicate the key findings of Kaup et al. (2007) using our own item set. This was an assumption underlying our study, and the replication suggests that the effects observed in the LMM analysis are not merely due to item-related artifacts.

Data Availability

Data and analysis script are available on OSF.

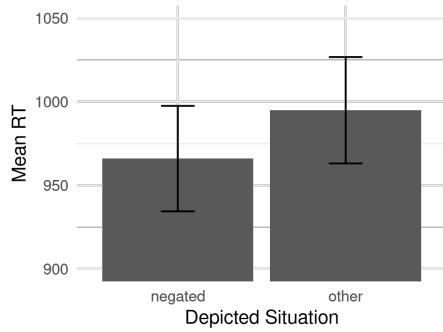


Figure 3: **Experiment 3** - Mean RTs $\pm$ standard errors.

Discussion

The results of our experiments highlight significant challenges in GPT-4o’s ability to process negation, aligning with patterns observed in human cognitive processing. In Experiment 1, we observed that GPT-4o produces more incorrect inferences when presented with negated sentences compared to affirmative ones. This pattern mirrors the well-documented “negation effect” in human cognition, where negation processing is associated with increased cognitive costs and higher error rates. Experiment 2 further corroborated these findings and provided additional insights into the nature of GPT-4o’s struggles with negation. The observation that false completions increase more substantially after negation than neutral completions suggests that the model’s difficulties extend beyond a simple underspecification due to failure to fully process negation. Instead, much like humans, GPT-4o disproportionately relies on incorrect affirmative representations even when faced with negation. Experiment 3 strengthens these conclusions by confirming that human participants exhibit similar difficulties with negation when tested with the same image-sentence pairs used in Experiment 1. The replication of Kaup et al. (2007) removes concerns that the observed errors of GPT-4o were due to dataset artifacts. Taken together, our results suggest that both GPT-4o and human participants exhibit response patterns consistent with an initial reliance on representations associated with the negated state, as predicted by the first stage of the two-step model of negation.

The findings from this study contribute several novel insights to existing research on negation processing in language models, particularly in the context of GPT-4o. The focus on GPT-4o is significant, as it moves beyond the limitations of earlier models like BERT and previous versions of GPT. One key contribution is the integration of visual and linguistic elements in examining negation processing. Previous research has primarily focused on text-based approaches, while this study explores how GPT-4o processes negation in the presence of visual stimuli, adding a new dimension to our understanding of AI’s capabilities in multimodal contexts.

Beyond this multimodal contribution, the study also focuses on qualitative patterns of negation processing. Rather than evaluating performance solely in terms of accuracy, we

examine how errors are structured and whether they resemble patterns predicted by cognitive models of negation, in particular the two-step model. However, an important limitation concerns the comparability of our paradigm to those used in psycholinguistic research. In Kaup et al. (2007), participants first process a sentence and are then presented with an image, allowing researchers to track how representations evolve over time. Evidence for two-step processing is based on the observation that responses are initially facilitated for images depicting the negated state, but later shift toward the actual state of affairs. In contrast, the present paradigm exposes the model to an image and requires it to complete a sentence. A closer procedural match to Kaup et al. (2007) is difficult, as GPT-4o does not provide reaction times or direct access to intermediate representations. For this reason, our experiments focus on the model’s final completions rather than on processing dynamics. Importantly, however, the use of the same image set and corresponding sentence materials allows us to test whether the model’s responses are structured around the same negated-state contrasts that drive human effects.

Accordingly, our results should not be taken as evidence that GPT-4o implements a two-step process. Rather, they show that the model’s responses are systematically biased toward completions associated with the affirmative counterpart, even in the presence of negation. One possible interpretation is that these responses are driven by associative or distributional regularities linking objects and contexts (e.g., cracked eggs and frying pans). In psycholinguistic accounts, early stages of processing are often assumed to be influenced by such activation-based dynamics before more context-sensitive integration occurs. From this perspective, GPT-4o’s behavior may resemble an overreliance on these associative patterns, without a mechanism to reliably override them when negation is present. However, our data do not allow us to determine whether similar underlying mechanisms are involved in human and model processing.

Overall, our results suggest that simply increasing model size or training data may not be sufficient to overcome the challenges posed by negation. The scope of our study is limited to a specific model and we did not directly compare models of different sizes or training scales. However, we base this claim on the observation that even newer generations of LLMs, despite increased model size and training data, continue to exhibit similar difficulties with negation. Instead, more targeted approaches that explicitly address the processes involved in negation comprehension may be necessary. This view resonates with Asher and Bhar (2024)’s approach, where negation is not merely another element in a latent representation but instead constrains the evolution of probability distributions over representations. Their work shows that treating negation as a standard token transformation in LLMs leads to *strong hallucinations*, as the models fail to preserve the logical constraints that negation imposes. While their solution focuses on enforcing logical consistency through probability distributions, a complementary direction

for improving negation handling in LLMs might come from incorporating insights from human negation processing.

Psycholinguistic research (e.g., Kaup et al., 2006, 2007) has shown that humans do not process negation instantaneously. Instead, negation comprehension might follow a two-stage process: an initial activation of the affirmative counterpart, followed by an integration step where negation modifies that representation. Our findings suggest that LLMs, however, tend to remain stuck at this first stage, primarily activating the affirmative meaning without fully integrating the negation. Moreover, psychological studies indicate that negation in human language processing is not treated merely as a logical operator but is heavily influenced by pragmatic considerations (Horn, 1989; Nieuwland and Kuperberg, 2008). Unlike formal semantic accounts, which treat negation as an operator over propositions - where a negated proposition is simply the complement of its affirmative counterpart - human comprehension of negation involves a more complex, context-dependent process. Experimental research suggests that when people encounter negation, they do not merely construct a truth-functional opposite; rather, they engage in pragmatic inferencing to determine what information is being conveyed or suppressed. For instance, negation often serves to correct presuppositions, highlight contrastive focus, or modify scalar implicatures (e.g., "not bad" implying "somewhat good" rather than outright logical negation). Alternatives to a negated entity tend not to be drawn indiscriminately from the entire complement set but are instead constrained to semantically similar concepts. That is, when a speaker negates an entity, listeners typically infer a substitute that maintains a close conceptual relationship to the original term rather than any arbitrary alternative (e.g., if one states, "This is not a dog", the inferred alternative is more likely to be a similar animal like a wolf or a cat; Capuano et al., 2021, 2023; Kruszewski et al., 2016). These effects are difficult to capture through purely logical approaches.

The results of this study suggest that the processing principles enabling humans to adequately capture negation and arrive at a representation of the actual state of affairs (i.e., the second step) could be beneficial for LLMs, which appear to have difficulties moving beyond affirmative interpretations (the first step). Future research should investigate whether explicitly integrating cognitive mechanisms could help mitigate the issues of the models with negation. On the other hand, our results raise questions into whether LLMs model only the surface level of linguistic comprehension rather than the full depth of human-like processing. In essence, they might behave like a fast, efficient processor of language's first-pass meanings but fall short when deeper semantic or context-sensitive revisions are needed. In conclusion, our study provides a nuanced understanding of how GPT-4o handles negation, underscoring similarities with human cognition.

Acknowledgments

Barbara Kaup received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SFB 1629 "Negation in Language and Beyond" (509468465) [Projects C2/C3]. Francesca Capuano and Lenny Döcker received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SFB1718 "Common Ground" (538184825) [Projects S and A5, respectively].

References

- Asher, N., & Bhar, S. (2024). *Strong hallucinations from negation and how to fix them*. arXiv: 2402.10543 [cs.CL].
- Capuano, F., Dudschig, C., Günther, F., & Kaup, B. (2021). Semantic similarity of alternatives fostered by conversational negation. *Cognitive Science*, 45(7), e13015. <https://doi.org/10.1111/cogs.13015>
- Capuano, F., Sorg, T., & Kaup, B. (2023). Activation levels of plausible alternatives in conversational negation. *Memory & Cognition*, 51(8), 1807–1818. <https://doi.org/10.3758/s13421-023-01434-2>
- Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological review*, 82(1), 45. <https://doi.org/10.1037/h0076248>
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive psychology*, 3(3), 472–517. [https://doi.org/10.1016/0010-0285\(72\)90019-9](https://doi.org/10.1016/0010-0285(72)90019-9)
- Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 34–48. https://doi.org/10.1162/tacl_a_00298
- García-Ferrero, I., Altuna, B., Alvez, J., Gonzalez-Dios, I., & Rigau, G. (2023). This is not a dataset: A large negation benchmark to challenge large language models. *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Horn, L. (1989). *A natural history of negation*. The University of Chicago Press.
- Jang, J., Ye, S., & Seo, M. (2023). Can large language models truly understand prompts? a case study with negated prompts. *Transfer learning for natural language processing workshop*, 203, 52–62. <https://proceedings.mlr.press/v203/jang23a.html>
- Johnson-Laird, P. (1983). *Mental models*. Cambridge MA Harvard University Press.
- Kassner, N., & Schütze, H. (2019). *Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly*. arXiv: 1911.03343 [cs.CL].
- Kaup, B., & Dudschig, C. (2020). Understanding negation: Issues in the processing of negation. In *The oxford handbook of negation*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198830528.013.33>

- Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7), 1033–1050. <https://doi.org/10.1016/j.pragma.2005.09.012>
- Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). Experiential simulations of negated text information. *Quarterly journal of experimental psychology*, 60(7), 976–990. <https://doi.org/10.1080/17470210600823512>
- Khandelwal, A., & Sawant, S. (2019). *Negbert: A transfer learning approach for negation detection and scope resolution*. arXiv: 1911.04211 [cs.CL].
- Kruszewski, G., Paperno, D., Bernardi, R., & Baroni, M. (2016). There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4), 637–660. https://doi.org/10.1162/COLI_a_00262
- Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12), 1213–1218. <https://doi.org/10.1111/j.1467-9280.2008.02226.x>
- Schütt, E., Dudschig, C., Bergen, B. K., & Kaup, B. (2023). Sentence-based mental simulations: Evidence from behavioral experiments using garden-path sentences. *Memory & Cognition*, 51(4), 952–965. <https://doi.org/10.3758/s13421-022-01367-2>
- Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, 11(2), 92–107. <https://doi.org/10.1080/17470215908416296>
- Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. *Grounding cognition: The role of perception and action in memory, language, and thinking*, 22, 224–245.