

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Alternative-Space Structure Shapes Negation Resolution in Multimodal Generation

Permalink

<https://escholarship.org/uc/item/8w57m9hb>

Journal

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

Authors

Capuano, Francesca

Berger, Elisa

Weber, Sophia Christine

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

Alternative-Space Structure Shapes Negation Resolution in Multimodal Generation

Francesca Capuano (francesca.capuano@uni-tuebingen.de), Elisa Berger, Sophia Weber & Barbara Kaup

Department of Psychology, University of Tübingen

Abstract

Negation in natural language is not interpreted as a simple logical complement but is resolved selectively, depending on how linguistic and conceptual structure constrain relevant alternatives. Psycholinguistic research suggests that negation facilitates consideration of an alternative only when a privileged counterstate is available, either through contextual restriction or graded prominence. We use text-to-image generation as a complementary method to probe this selectivity under forced representational commitment: when negated language must be rendered visually, a concrete state must be instantiated. Across two preregistered experiments, participants evaluated how well generated images matched affirmative and negated prompts. Experiment 1 shows that negation is resolved more successfully for predicates encoding conventionalized contrasts than for those with open alternative spaces. Experiment 2 shows that, within open spaces, negation is facilitated when a highly prominent alternative exists. Together, the results support an account in which binary advantages arise from how strongly a counterstate is privileged.

Keywords: Language Processing; Negation; Alternatives; Visual World; Multimodal Models; Stable Diffusion XL.

Introduction

Negation is a fundamental aspect of human language and cognition, enabling speakers to express denial, absence, and the violation of expectations (Givón, 1978; L. Horn, 1989; Wason, 1965). Despite its central role, negation poses well-documented processing challenges: negated statements are read more slowly, remembered less accurately, and verified with higher error rates than affirmative statements (Carpenter & Just, 1975; Clark & Chase, 1974; Just & Carpenter, 1971; Kaup & Dudschig, 2020; Wason, 1961). Classic evidence comes from sentence–picture verification tasks, where participants are slower and less accurate at verifying the truth of negated sentences (e.g., “The star is not above the plus”) than their affirmative counterparts (Carpenter & Just, 1975). These findings suggest that constructing and updating mental representations under negation places additional demands on comprehenders.

Artificial language systems exhibit parallel difficulties. Large language models (LLMs) struggle with interpreting and generating negated statements across tasks (Ettinger, 2020; García-Ferrero et al., 2023; Jang et al., 2023; Kassner & Schütze, 2019; Truong et al., 2023). For example, models may assign identical truth judgments to a sentence and its negation (e.g., endorsing both “Bills are commonly part of

birds” and “Bills are never part of birds”), effectively ignoring the negation operator. Such failures are striking given models’ otherwise strong linguistic performance, and suggest that difficulties with negation may reflect general properties of meaning construction rather than system-specific limitations. This convergence motivates a bidirectional research strategy: insights from human cognition can inform model evaluation, while systematic probing of models can generate hypotheses about mechanisms that may also operate in human processing. Indeed, instruction and dialogue fine-tuning improve model performance with negation and bring it closer to subtle human behavioral patterns (Capuano & Kaup, 2024; Jara-Ettinger, 2019; Truong et al., 2023), consistent with evidence that pragmatic reasoning and interaction play a key role in how humans acquire and use negation (L. Horn, 1989; Nordmeyer & Frank, 2014; Reuter et al., 2018; Schindele et al., 2008).

A critical feature of negation is that it is not interpreted as strict logical operator (L. Horn, 1989; Kruszewski et al., 2016). In contrast to formal logic, where negation denotes the exhaustive complement of a proposition, linguistic negation is shaped by pragmatic principles of relevance and informativeness (Grice, 1975; L. Horn, 1989; L. R. Horn, 1984). Speakers rarely use negation to describe all states of affairs in which a proposition fails to hold. Instead, negation is typically employed to reject a specific assumption or expectation that is salient in the discourse. For example, “That’s not a dog!” upon encountering a cat expresses a corrective stance toward the expectation that the animal was a dog, rather than asserting the set of all non-dog entities. Accordingly, negation is best understood as a pragmatically specialized, corrective device. It often makes alternative states of affairs relevant, but these alternatives are not determined by logical exhaustiveness; instead, their consideration depends on contextual factors such as salience, relevance and cognitive accessibility.

This perspective closely relates to work in alternative semantics and the study of linguistic focus, where meaning is characterized in terms of sets of contextually relevant alternatives (Rooth, 1992). In this tradition, interpretation depends not on enumerating all logically possible alternatives, but on how linguistic structure and discourse context restrict the set of candidates that are considered relevant. Psycholinguistic studies show that comprehenders rapidly activate such alternatives during online processing, including alternatives that are not explicitly mentioned, and that this activation is sen-

sitive to how strongly the context privileges particular candidates (Gotzner & Spalek, 2019; Gotzner et al., 2016; Repp & Spalek, 2021). A central question across these approaches is therefore not simply whether alternatives are available, but how constrained or permissive the relevant alternative set is, and how this affects interpretation. Converging evidence from electrophysiological studies supports this view: ERP responses to negation are modulated by contextual constraint and predictability, indicating that negation processing is facilitated when the input strongly constrains the expected alternative (Spychalska et al., 2025).

The visual world paradigm (VWP) has been a particularly informative method for studying negation in humans (Lüdtke et al., 2008; Orenes et al., 2014, 2016; Rück et al., 2021). By tracking eye movements as participants integrate linguistic input with visual context, VWP studies reveal how attentional patterns shift in response to negated statements. This paradigm can provide insight into when and how listeners consider alternative states of affairs in response to negation. For example, Orenes et al. (2014) manipulated the verbal context preceding a negated description while keeping the visual display constant. In a binary verbal context, which restricted the relevant alternatives to two possibilities (e.g., “The figure could be red or green”), listeners rapidly shifted their gaze toward the implied affirmative alternative upon hearing a negated sentence (e.g., “The figure is not red”). In contrast, in a multary verbal context, which left several alternatives contextually available, listeners tended to maintain visual attention on the negated referent itself. These findings suggest that negation processing is sensitive to how the linguistic context constrains the space of relevant alternatives.

Complementary evidence comes from work on automatic alternative activation showing that negation can trigger the activation of alternatives even without an explicitly closed set of options, but crucially, this activation is selective rather than exhaustive. In particular, alternatives are activated reliably only when one candidate is especially salient or dominant. Following Capuano et al. (2023), we refer to this property as *alternative prominence*, that is, the extent to which one alternative is strongly privileged within an otherwise open set of possibilities. For example, for a negated noun such as *not an onion*, comprehenders may be especially likely to consider *garlic*, because it is a highly prominent alternative in memory and world knowledge. By contrast, for a negated noun such as *not a chair*, no single alternative may stand out (e.g., table, sofa, stool, bench). Empirically, prominence can be operationalized using cloze tasks (“Das ist kein X, sondern ...”, Eng. “This is not an X, but rather ...”), where items are more prominent when one continuation is produced with particularly high probability. (Capuano et al., 2023) show that negation preferentially activates such highly dominant alternatives, whereas it fails to reliably activate any specific alternative when no dominant candidate stands out. This suggests that “binary-like” behavior may arise not only from the number of available alternatives, but from how strongly

linguistic material (and world knowledge) privileges one alternative over others.

Together, these findings converge on the idea that negation resolution depends on the availability of a privileged alternative. What remains less clear is whether these effects reflect paradigm-specific mechanisms or instantiate a more general principle: negation yields coherent alternative representations when the input sufficiently constrains the alternative space - either categorically, by limiting the number of relevant alternatives (e.g. red vs. green), or gradationally, by strongly privileging one alternative within an otherwise open set.

Multimodal text-to-image generation offers a complementary methodological perspective. Unlike comprehension tasks that allow underspecification, text-to-image models must externalize an interpretation: generating an image for a negated prompt requires committing to some concrete state of affairs. This forced representational commitment makes alternative selection directly observable in the output. While model behavior cannot be taken as a model of human cognition, it provides a useful testbed for probing when negation can be resolved into a coherent alternative and when it fails, thereby sharpening theoretical hypotheses about alternative selection under negation.

The present study uses text-to-image generation as a diagnostic tool to investigate how the structure of the alternative space affects negation resolution. In Experiment 1, we test whether predicate-internal contrast structure alone facilitates negation resolution. We compare predicates that conventionally encode a strong contrast (e.g., *not empty* $\rightarrow$ *non-empty/full*) with predicates that evoke open, weakly structured alternative spaces (e.g., *does not contain vegetables*), while not pre-specifying or visually constraining the set of possible alternatives compatible with the negation. (cf. “red vs. green” in Orenes et al., 2014). This allows us to ask whether an apparent “binary advantage” can arise from linguistic contrast structure itself, independent of explicit contextual enumeration. However, differences between binary-like and multinary-like predicates could reflect either a qualitative distinction between closed and open alternative sets, or a more graded sensitivity to how strongly a particular alternative is made salient by the linguistic and conceptual context. Experiment 2 therefore focuses within open alternative spaces, testing whether negation is resolved more successfully when a highly dominant alternative is available (Capuano et al., 2023).

Experiment 1

Experiment 1 tested whether negation is easier to externalize into a coherent visual counterstate when predicates encode a conventionalized contrast (binary) compared to when they evoke an open alternative space (multinary). Participants completed an online sentence-image matching task: on each trial they saw a sentence (affirmative or negated) together with an image generated from that prompt by Stable Diffusion XL, and rated how well the image matched the sentence on a 6-

point accuracy scale. The design crossed polarity (affirmative vs. negated) with predicate arity (binary vs. multinary) within participants.

Methods

The study was preregistered on AsPredicted. Data, materials, and analysis scripts are available on OSF.

Participants We targeted 100 participants; 92 were recruited. Following Brysbaert (2019), this sample is sufficient to detect medium-sized interactions. After exclusions (incomplete responses, underage, and fast responses $< 2s$; $n = 12$), the final sample comprised 80 participants (54 women, 25 men, 1 diverse; age: $M = 31.55$, $SD = 15.23$). All were German-proficient and provided informed consent.

Participants were recruited from psychology courses at the University of Tübingen and online; students received course credit, others participated voluntarily.

Materials To construct prompts that were comparable across polarity and predicate arity, we conducted a pre-study to control for scenario plausibility. This is important because scenario likelihood can affect how faithfully a text-to-image model depicts a prompt independently of negation or predicate structure (e.g., *The bicycle does not have inflated tires* is less typical than *The bicycle has inflated tires*). If negated scenarios are systematically less plausible, lower image-sentence match could reflect atypicality rather than negation processing.

We initially created 47 binary and 47 multinary prompts and paired them by topic to improve comparability across arity conditions (e.g., *The cat does not lie on the sofa* vs. *The cat's eyes are not open*). For the plausibility pre-study, 12 participants (SoSci Survey) rated the likelihood of the situations described by the 47 affirmative prompts on a 0–100% scale. We then used these ratings to select items such that both the affirmative and corresponding negated scenarios fell within a moderate plausibility range (35–65%). This yielded 18 binary and 18 multinary items, resulting in 72 experimental prompts total (36 affirmative, 36 negated).

Where possible, items were matched for overall scenario meaning, which introduced a trade-off: binary-like items were often adjectival, whereas multinary items were often verbal, so predicate type was not fully controlled. While this is a limitation for Experiment 1, predicate type was controlled in Experiment 2.

We used the open-source text-to-image model Stable Diffusion XL (stable-diffusion-xl-base-1.0; Stability AI), available via Hugging Face. Each prompt was submitted once to the model, generating one image per sentence. Prompts were provided in English; for the participant tasks, sentences were translated into German, as the available participant pool was predominantly German-speaking.

Procedure The experiment was administered online via SoSci Survey. After two practice trials, participants com-

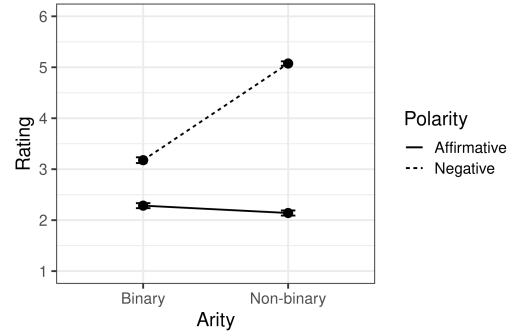


Figure 1: Experiment 1 - Mean accuracy ratings ($\pm se$). Participants evaluated how well the generated images matched the corresponding prompts using a 6-point scale, where 1 = very good and 6 = inadequate.

pleted all 72 experimental trials. On each trial, a generated image and its corresponding sentence were presented, and participants rated how accurately the image represented the sentence on a 6-point scale based on the German grading system (1 = very good, 6 = inadequate). Trials were presented in randomized order. The study was conducted in accordance with the ethical guidelines of the University of Tübingen.

Results

To evaluate the effects of negation and predicate arity on image accuracy, a two-way repeated-measures ANOVA was conducted with polarity (affirmative vs. negated) and arity (binary vs. multinary) as within-subject factors. Mean accuracy ratings for each condition are shown in Figure 1. The analysis revealed a significant main effect of polarity, $F(1, 79) = 1541$, $p < .001$, indicating a robust negation cost. The main effect of arity was also significant, $F(1, 79) = 559$, $p < .001$. Crucially, there was a significant interaction between polarity and arity, $F(1, 79) = 784.6$, $p < .001$.

As expected, accuracy ratings were worse for negated prompts than for affirmative prompts. Moreover, the significant interaction indicates that the decrement in accuracy due to negation was larger for multinary predicates than for binary predicates. This suggests that predicates with a binary contrast structure facilitate the interpretation of negated descriptions. See Figure 2 for example outputs.

Experiment 2

Experiment 2 isolates the role of alternative prominence within open alternative spaces. Using minimal copular sentences (e.g., “Das ist (k)ein X”), we generated one image per sentence with the text-to-image model integrated in ChatGPT 5.2 and asked participants to rate sentence-image match in the same online paradigm as Experiment 1. Polarity (affirmative vs. negated) and alternative prominence (prominent vs. non-prominent, based on cloze-dominance norms) were manipulated within participants, allowing us to test whether negation is rated as more successfully realized when a highly



(a) Binary, Affirmative:
The shopping basket is empty.



(b) Multinary, Affirmative:
The shopping basket contains vegetables.



(c) Binary, Negated:
The shopping basket is not empty.



(d) Multinary, Negated:
The shopping basket does not contain vegetables.

Figure 2: Experiment 1 - Example model outputs.

dominant alternative is available.

Methods

The study was preregistered on OSF. Data, materials, and analysis scripts are available on OSF.

Participants One hundred individuals participated; one was excluded due to incomplete responses. The preregistration specified the same exclusion criterion as Experiment 1, whereby trials with ratings under two seconds were to be excluded. However, because all sentences in this experiment were short and highly standardized (e.g., “Das ist (k)ein X”, Eng. “That is not an X”), rapid responses were expected and not informative about engagement. Applying the preregistered cutoff would have excluded the majority of participants ($n=72$). We therefore did not apply any threshold. Importantly, all analyses were additionally conducted using a more conservative cutoff of one second, and the results did not change qualitatively. The final sample comprised 99 participants (26 women, 73 men), aged 18-74 ($M = 33.06$, $SD = 11.75$). All participants were proficient in German and provided informed consent. Participants were recruited via Prolific and received £1 for approximately five minutes of participation.

Materials Materials were based on those used in Capuano et al. (2023), with one modification to the copular predicate. As in the original study, all items consisted of minimal sentences referring to concrete nouns and were classified as involving either prominent or non-prominent alternatives based on an independent cloze task. In that norming study, participants

completed prompts of the form “Das ist kein X, sondern” (Eng. “That is not an X, but rather”), and alternative prominence was operationalized as the relative frequency of the most common completion. From this normed item pool, we selected 36 items: the 18 items with the highest dominance scores were assigned to the prominent-alternative condition, and the 18 items with the lowest dominance scores to the non-prominent condition. Each item appeared in both affirmative and negated form, yielding 72 experimental sentences in total.

The only change relative to Capuano et al. (2023) concerned sentence form. Whereas the original study used “Dort ist (k)ein X” (Eng. “There is an/no X”), we used “Das ist (k)ein X”. This change was motivated by pilot testing with image-generation models. Stable Diffusion XL, used in Experiment 1, either failed to incorporate negation or produced severely distorted outputs for these materials. We therefore used ChatGPT 5.2 (accessed between 12. and 15. January 2026). For this model, sentences with “dort ist” tended to elicit images emphasizing spatial absence rather than alternative object identities, whereas “das ist” reliably elicited concrete object representations. Apart from this change in copular predicate and model, all materials and condition assignments were identical to those used in Capuano et al. (2023).

Procedure The procedure was identical to Experiment 1.

Results

To evaluate the effects of negation and alternative prominence on image accuracy, a two-way repeated-measures ANOVA was conducted with polarity (affirmative vs. negated) and prominence (prominent vs. non-prominent) as within-subject factors. Mean accuracy ratings for each condition are shown in Figure 3. The analysis revealed significant main effects of polarity, $F(1, 98) = 2137$, $p < .001$, and prominence, $F(1, 98) = 441.6$, $p < .001$, as well as a significant interaction between polarity and prominence, $F(1, 98) = 484.8$, $p < .001$. Overall, accuracy ratings were worse for negated sentences than for affirmative sentences, indicating a robust negation cost. In addition, sentences involving prominent alternatives were rated as more accurately represented than sentences involving non-prominent alternatives. Crucially, the significant interaction indicates that the effect of negation depended on alternative prominence: the decrement in accuracy due to negation was larger for non-prominent items than for prominent items.

This pattern shows that even within open alternative spaces, negation is resolved more successfully when a highly salient alternative is available. Importantly, this does not mean that the model consistently depicted the canonical alternative to the negated noun. In many cases, negated prompts elicited heterogeneous outcomes, including explicit negation marking (e.g., a prohibition sign over an alligator), depictions of alternatives that were not the most prominent according to human norms, hybrid or distorted objects, images emphasizing absence rather than replacement, or images in which the negated category was visually alluded to (e.g., via costumes) rather

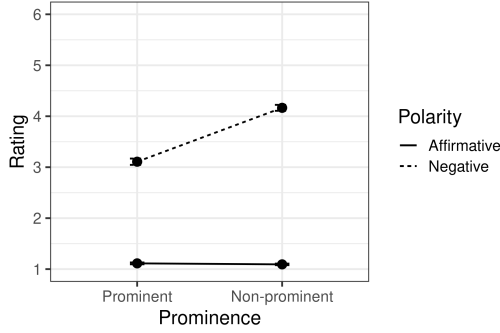


Figure 3: Experiment 2 - Mean accuracy ratings ($\pm se$).

than depicted directly (e.g., “kein Tiger” rendered as a cat wearing tiger-like features).

Nevertheless, the prominence manipulation reliably modulated accuracy ratings: negated sentences associated with highly dominant alternatives were judged as better matched than those with diffuse alternative distributions. This indicates that alternative prominence affects the model’s ability to produce evaluably coherent representations under negation, even when the output does not correspond to a single, well-defined substitute. In this sense, the effect of prominence does not seem tied to the explicit depiction of a specific alternative, but to the overall success with which negation can be externalized into a visually interpretable state of affairs. See Figure 4 for example outputs.

Discussion

The present work used text-to-image generation as a complementary methodological paradigm for investigating how negation is resolved when linguistic input does not uniquely specify a state of affairs. Psycholinguistic research has shown that negation is not interpreted as a strict logical complement, but is resolved selectively depending on the structure and salience of available alternatives. In visual-world studies, listeners shift attention toward an implied alternative primarily when contextual constraints restrict the alternative space (Orenes et al., 2014). Complementary evidence from automatic alternative activation suggests that, even without explicit contextual enumeration, negation reliably activates alternatives mainly when a particularly prominent candidate is available (Capuano et al., 2023). Building on these converging findings, we asked whether similar selectivity emerges when negation must be externalized into an explicit visual representation. It is important to clarify the level at which the present paradigm is intended to connect to human negation research. The relevant comparison is not between model image generation and an analogous offline image-production task in humans. Rather, the question is whether a system required to externalize negation shows the same sensitivity to counterstate availability that has been observed in human online processing.

In Experiment 1, participants rated how accurately images



(a) Prominent, Affirmative:
That is an onion.



(b) Non-prominent, Affirmative:
That is a chair.



(c) Prominent, Negated:
That is not an onion.



(d) Non-prominent, Negated:
That is not a chair.

Figure 4: Experiment 2 - Example model outputs.

generated by Stable Diffusion XL matched affirmative and negated prompts differing in predicate arity. We observed a robust negation cost and a polarity $\times$ arity interaction, with larger accuracy decrements for multinary than for binary predicates. This supports the view that negation is easier to resolve when linguistic structure provides a readily available counterstate. Binary predicates such as *empty* encode conventional contrasts that guide interpretation under negation, whereas multinary predicates such as *contain vegetables* evoke broader, less structured alternative spaces. A key contribution of Experiment 1 is that this asymmetry was implemented through predicate semantics rather than explicit enumeration of alternatives. Unlike Orenes et al. (2014), where verbal context explicitly restricts the alternative set, our manipulation varied the extent to which predicates themselves privileged a counterstate. Although both binary and multinary negated prompts are compatible with many possible world states, binary predicates provide a conventionalized contrast that can serve as a default resolution. This suggests that negation success depends not on the cardinality of the alternative space per se, but on how linguistic structure organizes that space and highlights a relevant counterstate.

Experiment 2 sharpened this interpretation by holding contrast structure constant and manipulating only alternative prominence. All items involved open, multinary alternative spaces; what varied was whether world knowledge strongly favored a single alternative. Using normed materials from

Capuano et al. (2023), we found a polarity $\times$ prominence interaction: negated prompts were rated as more accurately represented when a highly dominant alternative was available. Importantly, this effect did not depend on the consistent depiction of a canonical substitute. Instead, prominent items tended to yield outputs that were more coherent or interpretable overall, even when negation was conveyed through symbolic marking or near-miss representations. Interestingly, cases in which the model failed to resolve negation and instead depicted the negated referent bear a qualitative resemblance to patterns observed in human processing. In visual-world studies, listeners sometimes maintain attention on the negated state of affairs when the alternative space is weakly constrained (Orenes et al., 2014). While the underlying mechanisms might differ, this parallel suggests that failures to resolve negation may reflect similar constraints on alternative selection under conditions of low informativeness.

Together, the two experiments indicate that apparent “binary advantages” in negation do not arise from the number of alternatives per se, but from the extent to which linguistic or conceptual structure constrains interpretation. Such constraints can be categorical, as in predicates with conventionalized contrasts (Experiment 1), or graded, as in distributions of alternative prominence derived from world knowledge (Experiment 2). In both cases, the critical factor is whether negation can be resolved away from the affirmative default into some coherent counterstate.

Methodologically, the present work highlights the value of multimodal generation as a tool for studying meaning. The advantage of this paradigm is not that AI-generated images are intrinsically preferable to human-designed stimuli, but that text-to-image systems cannot leave negation underspecified: generating an image for a negated description requires committing to some concrete state of affairs, making the outcome of negation resolution observable and measurable via human judgments. While model behavior cannot be taken as direct evidence about human processing, this forced representational commitment can sharpen theoretical hypotheses about the role of contrast structure and alternative salience.

In light of this methodological perspective, the present results highlight a more specific unresolved issue in the interpretation of prominence effects in human negation research: what makes one alternative more available than others. Prior work has typically operationalized alternative dominance via cloze completion, which provides a useful behavioral index of prominence but does not explain why some concepts yield sharply peaked alternative distributions whereas others do not. For example, items such as *onion* may privilege a highly dominant alternative, whereas items such as *chair* do not; this difference could arise from factors including taxonomic neighborhood structure, habitual contrasts in experience, perceptual similarity, communicative conventions, or broader event and world knowledge. In this sense, *alternative prominence* is best understood not as a fully explanatory notion, but as a higher-level descriptive construct that must be decomposed

into more specific representational factors. The present data suggest one route for doing so. Qualitatively, negated prompts did not fail uniformly; instead, the model relied on different strategies, including depiction of plausible alternatives, symbolic negation marking, hybrid or near-miss representations, and depictions of absence rather than replacement. If these strategies vary systematically with predicate meaning and conceptual structure, their distribution may help identify what kinds of information support negation resolution when no explicit counterstate is given. More broadly, such regularities could generate hypotheses for human research by distinguishing cases in which negation is resolved through substitution of a salient alternative from cases in which it is supported by other resources, such as perceptual resemblance, symbolic conventions, or schematic representations of absence.

Several limitations should be noted. First, the dependent measure was a subjective accuracy rating, which may reflect multiple sources of mismatch, including visual quality. However, the consistent polarity effects and targeted interactions capture systematic differences in how negation is handled across conditions, rather than noise in image quality alone. Second, Experiment 2 employed a different image-generation system than Experiment 1 for practical reasons, limiting strict comparability across model architectures. However, the primary aim of the present work is not to characterize how different models implement negation, but to use them as methodological tools to probe how negation is resolved under forced representational commitment.

In sum, the present study shows that negation difficulties in multimodal generation are not uniform, but depend systematically on how strongly linguistic input constrains the space of relevant alternatives. Across a predicate-based manipulation (Experiment 1) and a prominence manipulation (Experiment 2), negation was more successfully resolved when linguistic or conceptual structure privileged a counterstate. These findings align with psycholinguistic accounts in which negation serves a corrective function and interpretation proceeds by constraining relevant alternatives rather than by computing a logical complement. More broadly, the results illustrate how multimodal generative systems can function as hypothesis-generating tools for theories of meaning representation, helping to clarify which aspects of alternative structure facilitate resolution and where existing human paradigms leave open how negation is ultimately represented.

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 received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SFB1718 “Common Ground” (538184825) [Project S].

References

- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 16. <https://doi.org/10.5334/joc.72>
- Capuano, F., & Kaup, B. (2024). Pragmatic reasoning in gpt models: Replication of a subtle negation effect. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. <https://escholarship.org/uc/item/22q5920s>
- 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), 45–73. <https://doi.org/10.1037/h0076248>
- Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory & Cognition*, 2(1), 101–111. <https://doi.org/10.3758/bf03197499>
- 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., Álvarez, J., Gonzalez-Dios, I., & Rigau, G. (2023). This is not a dataset: A large negation benchmark to challenge large language models. *arXiv preprint arXiv:2310.15941*. <https://doi.org/10.48550/arXiv.2310.15941>
- Givón, T. (1978). Negation in language: Pragmatics, function, ontology. In *Pragmatics* (pp. 69–112). Brill. https://doi.org/10.1163/9789004368873_005
- Gotzner, N., & Spalek, K. (2019). The life and times of focus alternatives: Tracing the activation of alternatives to a focused constituent in language comprehension. *Language and Linguistics Compass*, 13(2), e12310. <https://doi.org/10.1111/lnc3.12310>
- Gotzner, N., Wartenburger, I., & Spalek, K. (2016). The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, 8(1), 59–95. <https://doi.org/10.1017/langcog.2015.25>
- Grice, H. P. (1975). Logic and conversation. In *Speech Acts* (pp. 41–58). Brill. https://doi.org/10.1163/9789004368811_003
- Horn, L. (1989). *A natural history of negation*. The University of Chicago Press. [https://doi.org/10.1016/0024-3841\(90\)90064-R](https://doi.org/10.1016/0024-3841(90)90064-R)
- Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Georgetown University Press. <https://hdl.handle.net/10822/555477>
- 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*, 52–62. <https://doi.org/10.48550/arXiv.2209.12711>
- Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29, 105–110. <https://doi.org/10.1016/j.cobeha.2019.04.010>
- Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10(3), 244–253. [https://doi.org/10.1016/s0022-5371\(71\)80051-8](https://doi.org/10.1016/s0022-5371(71)80051-8)
- Kassner, N., & Schütze, H. (2019). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*. <https://doi.org/10.18653/v1/2020.acl-main.698>
- Kaup, B., & Dudschig, C. (2020). Understanding negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), *The oxford handbook of negation* (pp. 635–655). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198830528.001.0001>
- 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
- Lüdtke, J., Friedrich, C. K., De Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence–picture verification paradigm. *Journal of Cognitive Neuroscience*, 20(8), 1355–1370. <https://doi.org/10.1162/jocn.2008.20093>
- Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children’s comprehension of negation. *Journal of Memory and Language*, 77, 25–39. <https://doi.org/10.1016/j.jml.2014.08.002>
- Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of memory and language*, 74, 36–45. <https://doi.org/10.1016/j.jml.2014.04.001>
- Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *Quarterly Journal of Experimental Psychology*, 69(6), 1082–1092. <https://doi.org/10.1080/17470218.2015.1063675>
- Repp, S., & Spalek, K. (2021). The role of alternatives in language. *Frontiers in Communication*, 6, 682009. <https://doi.org/10.3389/978-2-88966-840-3>
- Reuter, T., Feiman, R., & Snedeker, J. (2018). Getting to no: Pragmatic and semantic factors in two- and three-year-olds’ understanding of negation. *Child development*, 89(4), e364–e381. <https://doi.org/10.1111/cdev.12858>
- Rooth, M. (1992). A theory of focus interpretation. *Natural language semantics*, 1(1), 75–116. <https://doi.org/10.1007/bf02342617>
- Rück, F., Dudschig, C., Mackenzie, I. G., Vogt, A., Leuthold, H., & Kaup, B. (2021). The role of predictability during negation processing in truth-value judgment tasks. *Journal*

- of psycholinguistic research, 50, 1437–1459. <https://doi.org/10.1007/s10936-021-09804-0>
- Schindele, R., Lüdtke, J., & Kaup, B. (2008). Comprehending negation: A study with adults diagnosed with high functioning autism or asperger's syndrome. *Intercultural Pragmatics*, 5(4). <https://doi.org/10.1515/iprg.2008.021>
- Spychalska, M., Haase, V., & Werning, M. (2025). To predict or not to predict: The role of context constraint and truth-value in negation processing. *Neuropsychologia*, 216, 109167. <https://doi.org/10.1016/j.neuropsychologia.2025.109167>
- Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*. <https://doi.org/10.18653/v1/2023.starsem-1.10>
- Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2), 133–142. [10.1111/j.2044-8295.1961.tb00775.x](https://doi.org/10.1111/j.2044-8295.1961.tb00775.x)
- Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4(1), 7–11. [https://doi.org/10.1016/s0022-5371\(65\)80060-3](https://doi.org/10.1016/s0022-5371(65)80060-3)