Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Knowing the Scale, Missing the Negation: Language Models and Corrective Negation by Excess in English, Spanish, and French

Maikel Leyva

maikel.leyvav@ug.edu.ec

Universidad de Guayaquil https://orcid.org/0000-0001-7911-5879

Alexis Matheu Pérez

Universidad Bernardo O'Higgins https://orcid.org/0000-0002-7859-9039

Bastián Gutiérrez Vargas

Universidad Bernardo O'Higgins https://orcid.org/0000-0002-0056-6624

Research Article

Keywords: negation, metalinguistic negation, scalar adjectives, large language models, multilingual evaluation, corrective conjunctions

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11167683/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

In “it is not warm but boiling,” the negation does not deny that the soup is warm; it rejects “warm” as too weak. We test whether large language models understand this corrective negation by excess. From three resources of intensity-ordered adjective scales in English, Spanish, and French, we built 427 adjective pairs and queried five models under several conditions. A model–pair unit was analysed as *known* when the model judged the stronger adjective a stronger degree of the weaker one and rejected the reverse. On known units, models accepted the affirmative entailment (“it is dark; is it at least dim?”) in 93.4% of cases, but failed the corresponding corrective negation (“it is not dim but dark”) on 452 units where they passed the affirmative, against 6 in the opposite direction. When the stronger term was the same adjective intensified (“it is not dim but very dim”), accuracy remained high (94.4% vs. 39.8% for the lexical corrective on units that also passed the affirmative). The pattern held in all three languages, in all five models, on a subset of 247 pairs screened by the first author for scale validity, and on 48 pairs known to every model. A reversed question showed that the failures on the lexical corrective mix a literal reading of the negation with answers of “no” to both questions. In English, the explicit corrective *but rather* improved accuracy over *but* on the same items (49.3% vs. 36.1%). Models thus have the scalar entailment available but often do not apply it under corrective negation, and explicit marking of the correction improves accuracy.

1. Introduction

When a speaker says that a film was “not good but excellent,” the negation rejects “good” as an understatement, and the film is taken to be good and more than good. Horn (1985) called this use metalinguistic negation, a device for objecting to a previous or possible utterance on any grounds, including its scalar strength. Some languages mark the corrective use lexically. Spanish *sino* and German *sondern* introduce only corrections after a negation, while Spanish *pero*, English *but*, and French *mais* can also introduce contrast or concession (Anscombre & Ducrot, 1977); French *mais*, in particular, has both a corrective and an argumentative use.

Corrective negation by excess makes two demands on a language model. The model must know that the second adjective entails the first, and it must use that knowledge to override the literal reading of “not X.” Language models encode a good deal about scalar adjectives: they rank adjectives by intensity (Garí Soler & Apidianaki, 2020, 2021) and represent lexical properties of scales, though less their pragmatics (Lin et al., 2024). Negation is a documented weak point (Ettinger, 2020; Kassner & Schütze, 2020; Hossain et al., 2020; Hosseini et al., 2021; Truong et al., 2023; She et al., 2023). The interaction between the two has not been evaluated.

We ask whether models understand corrective negation by excess and, when they fail, whether the failure lies in their knowledge of the scale, in the scalar entailment, or in the corrective negation. For each adjective pair, a model is tested on the ordering and on the affirmative entailment before its reading of corrective negation is assessed. We compare plain negation, corrective negation with a different

stronger adjective, and corrective negation with the same adjective intensified. A reversed question probes response bias, and an English contrast between *but* and *but rather* varies only the corrective marker.

2. Background

Corrective and metalinguistic negation. Horn (1985, 1989) distinguished descriptive negation from metalinguistic negation, which objects to how something is expressed, including its implicatures; “she is not happy, she is ecstatic” objects to “merely happy.” Carston (1996) analysed such cases as negation applied to echoically used content, and Geurts (1998) as one of several mechanisms of denial. On these accounts, “not warm but boiling” entails that the subject is at least warm. We use that entailment as the criterion of a correct reading. Anscombre and Ducrot (1977) distinguished a corrective use of French *mais*, corresponding to Spanish *sino*, from an argumentative use corresponding to *pero*.

Scalar adjectives. Adjectives such as *warm*, *hot*, and *boiling* lie on a common scale ordered by intensity (Kennedy & McNally, 2005). We use three resources of gold-standard orderings. DEMELO (de Melo & Bansal, 2013) contains adjective sets extracted from WordNet “dumbbell” structures and divided into half-scales using pattern-based corpus evidence; its adjective pairs were manually annotated for intensity relations. CROWD (Cocos et al., 2018) and WILKINSON (Wilkinson & Oates, 2016) were built and ranked by crowd workers (Garí Soler & Apidianaki, 2020, Section 3). CROWD is available only in English. Garí Soler and Apidianaki (2020, 2021) used these resources to probe language models and released Spanish and French versions. Scales also differ in how readily they trigger scalar inferences (van Tiel et al., 2016; Gotzner et al., 2018).

Negation and inference in language models. Models often give similar predictions for negated and non-negated prompts (Ettinger, 2020; Kassner & Schütze, 2020). Natural language inference benchmarks contain few cases where negation matters (Hossain et al., 2020), and targeted benchmarks find weak negation reasoning (Geiger et al., 2020; She et al., 2023; Truong et al., 2023). Jeretic et al. (2020) evaluated scalar implicature in inference models, and McCoy et al. (2019) showed that such models can rely on lexical-overlap heuristics. Hu and Levy (2023) caution that prompted answers may understate what models represent. None of these studies addresses corrective negation by excess.

3. Method

3.1 Pairs

We used the English DEMELO, CROWD, and WILKINSON scales and the Spanish and French versions of DEMELO and WILKINSON. From each scale we took ordered pairs (X, Y) with Y ranked above X, kept single-word adjectives and at most three pairs per scale, and sampled 150 pairs per language with a fixed seed. We excluded 23 ungrammatical or invalid pairs (quantifiers, ordinals and multiplicatives, and

pairs whose antonym coincided with X), leaving 427 pairs (142 English, 139 Spanish, 146 French; 321 from DEMELO, 55 from WILKINSON, 51 from CROWD).

Because the resources contain pairs that are debatable as intensity scales, the first author also screened all 427 pairs and set aside 180. The screening was carried out after the main results had been computed, but without consulting any model's answers to individual pairs. Pairs were set aside in five categories: non-gradable or classificatory terms such as *nationwide/worldwide* (16); comparatives, superlatives including Spanish *-ísimo* forms, and ordinals, such as *high/higher* or *grande/grandísimo* (12); opposite poles or reversed pairs such as *chaud/froid* (4); pairs in which the second adjective names a different property or a different sense of the first, such as *cool/hostile* or *good/real* (117); and near-synonyms such as *closed/shut* (31). The screened set contains 247 pairs (79 English, 88 Spanish, 80 French). Because the screening reflects one annotator's judgement, every result is reported on both the full and the screened set, and the screening list is released with the data.

3.2 Conditions

For each pair, prompts were built without a subject noun:

- **A (ordering):** "Is 'Y' a stronger degree of 'X' (the same property, but more intense)? Answer only Yes or No." (correct: yes)
- **Arev:** the same with X and Y exchanged. (correct: no)
- **P (affirmative):** "It is Y." (correct: yes)
- **N (plain negation):** "It is not X." (correct: no)
- **E (lexical corrective):** "It is not X but Y." (correct: yes)
- **S (same-word corrective):** "It is not X but very X." (correct: yes)
- **D (downward corrective):** "It is not X but [antonym]." (correct: no)
- **E-rather (English only):** "It is not X but rather Y." (correct: yes)

Conditions P to E-rather were followed by "Based only on this sentence, is it at least 'X' (that is, X or more than X)? Answer only Yes or No." For N, E, and S we also asked "…is it less than 'X' (below the level X)?", with correct answers yes, no, and no. Spanish items used *No es X, sino Y* and *muy X*; French items used *Ce n'est pas X, mais Y* and *très X*. We also collected Spanish *No es X, pero Y*. Because *pero* after a negation is standardly concessive ("although not X, Y"), a "no" answer to the at-least question is defensible there, and we report that condition only descriptively.

3.3 Models and procedure

We queried five models through OpenRouter: GPT-4o-mini, GPT-5.4-nano, GPT-5.4-mini, Gemini 3.1 Flash-Lite, and Qwen 3.7 Flash, each in a new conversation, at temperature 0, with low reasoning effort where supported. Answers were parsed for *yes/no*, *sí/no*, or *oui/non*. We obtained 22,317 usable responses.

Llama 3.1 8B was also queried and excluded after inspection of its results: it passed the ordering and affirmative tests on 6.3% of English, 20.1% of Spanish, and none of the French pairs (37 known+P units in total), answered plain negation correctly on 48.6% of those units, and answered “yes” to 53% of all items. On its 37 units, accuracy was 56.8% on E and 78.4% on S. Mistral Small was dropped because rate limits allowed fewer than 2% of its queries to complete.

Several prompts are shared by different pairs (e.g., “It is not dim.” for every pair with X = *dim*), and some prompts were queried more than once. Of 1,780 prompts with more than one answer, 105 (5.9%) received inconsistent answers despite temperature 0. For each model and prompt we used the majority answer, dropped the 85 ties, and assigned the answer to every pair that uses the prompt. Resolving ties by the first answer instead of dropping them changed no paired count reported below by more than two units.

3.4 Filtering and analysis

A model–pair unit is **known** if the model answered A and Arev correctly, and **known+P** if it also answered P correctly. The comparison between P and E is made on known units, so that it is not guaranteed by the filter. Comparisons that require the affirmative entailment (E vs. S, reversed question, E-rather) are made on known+P units. Paired comparisons use exact McNemar tests, with Holm adjustment over the confirmatory family (P vs. E; E vs. S overall and per language; E-rather vs. E). Adjusted p-values below .001 are reported as $p < .001$. Logistic models fitted by generalised estimating equations, clustered by pair, estimate condition and language effects with model as a covariate; we report odds ratios with 95% confidence intervals.

4. Results

Unless stated otherwise, figures refer to the full set of 427 pairs; the corresponding figures for the screened set are given in parentheses.

4.1 Coverage

Of 2,135 model–pair units, 804 were known and 751 known + P (screened: 645 and 614). Models endorsed the resources’ orderings on well under half of the pairs; the analyses below are restricted to the pairs each model endorsed. Rates of “yes” across all responses ranged from 34% to 44%, below the proportion of items whose correct answer was yes (57%); the models therefore leaned towards “no.” This lean favours conditions whose correct answer is “no,” such as N and D, and works against E, so the high accuracy on N and D should not be read as evidence of full comprehension of negation.

4.2 Corrective negation breaks an entailment the models accept

Table 1 shows accuracy by condition. On known units, models accepted the affirmative entailment in 93.4% of cases and answered plain negation correctly in 99.2%, but answered the lexical corrective correctly in only 37.9%.

Table 1
Accuracy by condition (question: “is it at least X?”)

Condition	Example	All units	Known	Known + P
P	It is dark.	76.3%	93.4%	100%
N	It is not dim.	98.9%	99.2%	99.2%
E	It is not dim but dark.	21.8%	37.9%	39.8%
S	It is not dim but very dim.	92.6%	94.6%	94.4%
D	It is not dim but [antonym].	93.0%	93.1%	93.1%

On known units, models answered P correctly and E wrongly in 452 cases, and E correctly and P wrongly in 6 (adjusted $p < .001$; screened: 349 vs. 3). The asymmetry held in each language (English 188 vs. 1, Spanish 132 vs. 2, French 132 vs. 3).

The same-word corrective S shares the construction with E and differs only in whether the stronger term is a different word. On known + P units, S was right and E wrong in 414 cases, against 5 (adjusted $p < .001$; screened: 324 vs. 4), in every language and every model (Table 2). With condition, language, and model as predictors, the odds ratio for E relative to S was 0.024 (95% CI 0.016–0.036; screened 0.022, 0.014–0.035).

Table 2
Known + P units: accuracy by model

Model	P	N	E	S
Gemini 3.1 Flash-Lite	100%	100%	41.3%	97.4%
GPT-4o-mini	100%	100%	14.9%	82.2%
GPT-5.4-mini	100%	99.1%	67.0%	99.1%
GPT-5.4-nano	100%	99.2%	54.0%	98.6%
Qwen 3.7 Flash	100%	97.5%	35.4%	98.2%

To rule out that these differences reflect different item subsets per model, we examined the 48 pairs that were known + P for all five models (screened: 46). On these common items, accuracy on E ranged from 33.3% (GPT-4o-mini) to 77.1% (GPT-5.4-mini), while S was at ceiling for four models and 66.7% for GPT-4o-mini. GPT-5.4-mini and GPT-5.4-nano were close (77.1% and 75.0%), so the data do not support a clear ranking among the stronger models.

4.3 What the errors are

The reversed question distinguishes kinds of error. For “it is not dim but dark,” a coherent correct reading answers “yes” to “at least dim?” and “no” to “less than dim?”; a literal reading of the negation answers “no” and “yes”; answering “no” to both is incoherent. Accuracy on the reversed question for E was 68.4%, against 39.8% on the at-least question. On 751 known + P units, 252 (33.6%) were coherent and correct, 190 (25.3%) followed the literal reading, 262 (34.9%) answered “no” to both, and 47 (6.3%) answered “yes” to both (screened set, 614 units: 221, 145, 204, and 44). For the same-word corrective, 669 of 735 units (91.0%) were coherent, and the reversed question for plain negation was answered correctly on 74.5% of units, so a blanket tendency to say “no” after any negation does not explain the pattern. The “no to both” answers are specific to the lexical corrective. They are compatible with a reading on which the sentence says that Y holds instead of X, committing to neither degree of X, and with a default “no” in that construction; the data do not decide between these.

4.4 Explicit marking of the correction

In English, adding *rather* keeps every other word constant and marks the construction as corrective. *Rather* can also mean “fairly,” but in the frame “not X but rather Y” the corrective reading is the natural one. On known + P units, *but rather* was answered correctly in 49.3% of cases and *but* in 36.1% (55 units favoured *but rather*, 16 favoured *but*; adjusted $p < .001$; screened: 51.7% vs. 39.2%, 46 vs. 16). The direction held in four of the five models; GPT-4o-mini stayed near floor in both.

Descriptively, accuracy on the lexical corrective was higher in Spanish (*sino*) than in English (*but*) and French (*mais*) (Table 3). The three languages use different, partly translated items, so we do not attribute this difference to the conjunction. For Spanish *pero*, accuracy was 27.1%; as noted in Section 3.2, this condition is not a valid test of corrective reading.

Table 3
Known + P units: accuracy on the lexical corrective E by language and model

Model	English (*but*)	Spanish (*sino*)	French (*mais*)
Gemini 3.1 Flash-Lite	37.1%	54.4%	27.8%
GPT-4o-mini	9.9%	27.9%	6.1%
GPT-5.4-mini	61.0%	70.2%	70.8%
GPT-5.4-nano	44.6%	73.5%	41.2%
Qwen 3.7 Flash	40.6%	36.4%	26.7%

5. Discussion

Availability and use of the entailment. On the units analysed, a model has stated that Y is a stronger degree of X and has accepted that "it is Y" implies "it is at least X." Adding "not X but" before Y leads it, in most cases, to withhold that implication. The same-word corrective shows that the construction itself is not the obstacle, although its success may be helped by lexical overlap between X and "very X" (McCoy et al., 2019). The difficulty arises when the model must connect two different words inside a corrective negation.

Kinds of error. One quarter of the units follow a literal reading of the negation. This is compatible with Horn's (1989) account, in which a descriptive reading is computed first and revised when the correction arrives, with the revision missing here. A larger share answer "no" to both questions, which is specific to the lexical corrective and compatible with a reading on which Y holds instead of X. Our data are behavioural and do not establish the mechanism.

Explicit marking. In English, making the correction explicit with *rather* raised accuracy by 13 points on identical items. This fits the distinction between corrective and argumentative adversatives (Anscombre & Ducrot, 1977) and suggests that models are sensitive to how clearly a correction is signalled. Even with explicit marking, about half of the units failed. A clean test of Spanish *sino* would compare it with another well-formed corrective construction rather than with concessive *pero*.

Consequences. Evaluation suites for negation should include corrective negation by excess and should test whether models apply scale knowledge, not only whether they have it. Systems that summarise reviews or evaluations may misread strong praise ("not good but outstanding") and strong criticism ("not bad but disastrous").

6. Limitations

We did not collect human judgements on the comprehension items and make no claim about human performance. Scale orderings come from existing annotated resources whose pairs are sometimes debatable as intensity scales, and the screening of 180 pairs reflects one annotator and was done after the main results were known, though without consulting per-pair model answers; all results are therefore reported on both the full and the screened set. The exclusion of Llama 3.1 8B was decided after inspecting its results. The filters retain different pairs for each model, which we address with a common-item analysis on 48 pairs. Prompts use one template per condition, no subject noun, a yes/no format, and temperature 0; prompted answers may understate model knowledge (Hu & Levy, 2023). Repeated prompts occasionally received inconsistent answers at temperature 0 (5.9%), which we resolved by majority vote. The Spanish and French scales are translations, French *mais* is itself ambiguous between corrective and argumentative uses, and cross-language differences are descriptive. Five models were analysed, and model versions change over time.

7. Conclusion

Language models accept that “it is dark” implies “it is at least dim,” yet on most of the same pairs they do not conclude that something “not dim but dark” is at least dim. On units where the model knew the scale and accepted the affirmative entailment, accuracy was 39.8% on the lexical corrective and 94.4% on the same-word corrective (“not dim but very dim”). The failure appeared in five models and three languages, on a full and a screened set of pairs, and on items known to every model. It shrank when the correction was marked explicitly: in English, *but rather* gave 49.3% against 36.1% for *but* on identical items. What the models lack, then, is not knowledge of the scale but its use when a negation rejects the weaker term as too weak. We recommend that evaluations of negation include this construction and separate what models know about scales from whether they apply that knowledge under negation. Applications that summarise reviews, complaints, or teaching evaluations should be tested on such sentences, since they can read strong praise (“not good but outstanding”) as criticism.

Several lines of work follow. First, a human baseline on the same items would show how far the models depart from native speakers, who may also hesitate on debatable pairs. Second, the Spanish *sino* should be compared with another well-formed corrective construction, such as *sino más bien*, rather than with concessive *pero*, and the design can be extended to languages with a dedicated corrective conjunction, such as German *sondern*. Third, the downward direction (“not bad but terrible”) and constructions without an explicit conjunction (“not good; excellent”) deserve the same paired treatment. Fourth, varying the prompt template, adding subject nouns and context, and using free-text or graded answers instead of yes/no would test whether the effect depends on the format. Finally, probing internal representations could separate the two kinds of error we observed, a literal reading of the negation and a reading on which Y replaces X, and test whether fine-tuning or instructions that make the correction explicit close the gap.

Declarations

Data and code availability

Items, raw responses, the screening list, and analysis scripts will be deposited in a public repository upon acceptance.

References

1. Anscombre, J.-C., & Ducrot, O. (1977). Deux *mais* en français? *Lingua*, *43*(1), 23–40. https://doi.org/10.1016/0024-3841(77)90046-8
2. Carston, R. (1996). Metalinguistic negation and echoic use. *Journal of Pragmatics*, *25*(3), 309–330. https://doi.org/10.1016/0378-2166(94)00109-X
3. Cocos, A., Wharton, S., Pavlick, E., Apidianaki, M., & Callison-Burch, C. (2018). Learning scalar adjective intensity from paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1752–1762). https://doi.org/10.18653/v1/D18-1202

4. de Melo, G., & Bansal, M. (2013). Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, *1*, 279–290. https://doi.org/10.1162/tacl_a_00227
5. 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
6. Garí Soler, A., & Apidianaki, M. (2020). BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7371–7385). https://doi.org/10.18653/v1/2020.emnlp-main.598
7. Garí Soler, A., & Apidianaki, M. (2021). Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4653–4660). https://doi.org/10.18653/v1/2021.naacl-main.370
8. Geiger, A., Richardson, K., & Potts, C. (2020). Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 163–173). https://doi.org/10.18653/v1/2020.blackboxnlp-1.16
9. Geurts, B. (1998). The mechanisms of denial. *Language*, *74*(2), 274–307. https://doi.org/10.2307/417868
10. Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*, 1659. https://doi.org/10.3389/fpsyg.2018.01659
11. Horn, L. R. (1985). Metalinguistic negation and pragmatic ambiguity. *Language*, *61*(1), 121–174. https://doi.org/10.2307/413423
12. Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press.
13. Hossain, M. M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., & Blanco, E. (2020). An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 9106–9118). https://doi.org/10.18653/v1/2020.emnlp-main.732
14. Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordoni, A., & Courville, A. (2021). Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1301–1312). https://doi.org/10.18653/v1/2021.naacl-main.102
15. Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5040–5060). https://doi.org/10.18653/v1/2023.emnlp-main.306
16. Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics (pp. 8690–8705). https://doi.org/10.18653/v1/2020.acl-main.768

17. Kassner, N., & Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7811–7818). https://doi.org/10.18653/v1/2020.acl-main.698
18. Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381. https://doi.org/10.1353/lan.2005.0071
19. Lin, F., Altshuler, D., & Pierrehumbert, J. B. (2024). Probing large language models for scalar adjective lexical semantics and scalar diversity pragmatics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 13033–13049). https://doi.org/10.63317/28thzn779nt3
20. McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448). https://doi.org/10.18653/v1/P19-1334
21. She, J. S., Potts, C., Bowman, S. R., & Geiger, A. (2023). ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1803–1821). https://doi.org/10.18653/v1/2023.acl-short.154
22. Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)* (pp. 101–114). https://doi.org/10.18653/v1/2023.starsem-1.10
23. van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175. https://doi.org/10.1093/jos/ffu017
24. Wilkinson, B., & Oates, T. (2016). A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2669–2675). https://doi.org/10.63317/5q95e9epid2n