It is difficult, but not impossible: Measuring Scalar Activation in Language Models

David Arps and **Yulia Zinova**
Heinrich Heine University Düsseldorf
david.arps@hhu.de, zinova@hhu.de

Abstract

We explore the behaviour of language models on adjectival scales by analyzing activation changes when prompted with related and unrelated adjectives. We find evidence for scale activation, which aligns with results from human priming experiments.[1]

1 Introduction

Scalar diversity has been extensively studied in experimental setups with human participants when testing implicature endorsement rates (Van Tiel et al., 2016; Sun et al., 2018; Gotzner et al., 2018; Ronai and Xiang, 2022). Priming experiments (e.g. Lacina and Gotzner, 2024) explore the link between implicature computation and lexical priming. They find that priming with a weak scalemate leads to faster recognition of the strong scalemate.

Hu et al. (2023) show that pragmatic inference tasks pose great challenges for language models (LMs). Nizamani et al. (2024) show that DeBERTa models perform poorly on scalar implicatures, even after fine-tuning. As the availability of alternatives is considered to be the basis of implicature computation (Gotzner and Romoli, 2022), we analyze the activation of scalar adjectives in the LM.

2 Experimental Setup

Activation of strong adjectives In our first experiment, we follow the design used in Lacina and Gotzner (2024) and Ronai and Xiang (2023). In human priming experiments, participants were presented with sentences that carry either a related (1-a) or an unrelated (1-c) adjective. After that, participants were asked to perform a lexical decision task and their reaction times were recorded. Both Ronai and Xiang (2023) and Lacina and Gotzner (2024) found that participants recognized stronger adjectives as existent words faster when the preceding sentence contained the weak scalar item.

[1]We will release our code upon publication.

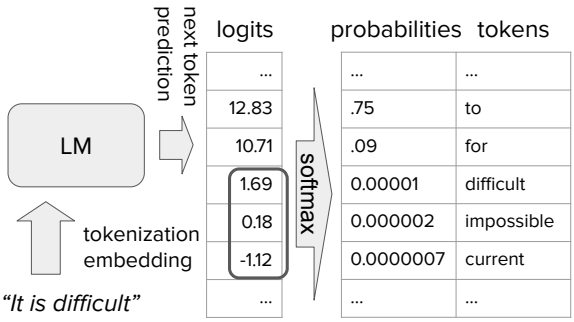

Figure 1: We collect next word prediction logits from the LM to measure activation of scalar concepts.

(1) a. It is difficult. (WEAK)
b. It is impossible. (STRONG)
c. It is current. (UNRELATED)

To test whether a similar effect can be observed within a LM, we collect activations for strong adjectives after either a weak or an unrelated adjective has been processed. The hypothesis that corresponds to human behaviour is that the activation of the strong adjective should be higher after the model processes a weak adjective, in comparison with processing an unrelated adjective.

Activation of weak adjectives We invert the prime/target adjectives from the previous setup and collect activations of the weak adjectives given either a related (1-b) or an unrelated (1-c) prompt.

Activation difference We use both setups above to check whether LM behaviour aligns with the results of De Carvalho et al. (2016) for humans, who found that weak terms activate the respective strong ones more than strong terms activate weak ones (in French).

Activation of unrelated adjectives As a control condition, we collect activations of unrelated adjectives after prompts with weak and strong adjectives.

Activation without context Ronai and Xiang (2023) did not find evidence for priming when par-

activation condition	Lacina and Gotzner (2024) of strong weak, unr.	of weak strong, unr.	diff.	of unr. strong, weak	No Context of strong weak, unr.	of weak strong, unr.	diff.	of unr. strong, weak
125M	****	****	n.s.	n.s.	****	****	n.s.	n.s.
350M	****	***	n.s.	*	****	*	*	*
1.3B	****	****	n.s.	n.s.	****	****	n.s.	n.s.
2.7B	****	****	n.s.	n.s.	****	****	n.s.	n.s.
6.7B	****	****	n.s.	n.s.	****	****	n.s.	n.s.

Table 1: Significance test results, where * stands for $p < 0.05$, *** for $p < 0.001$ and **** for $p < 0.0001$.

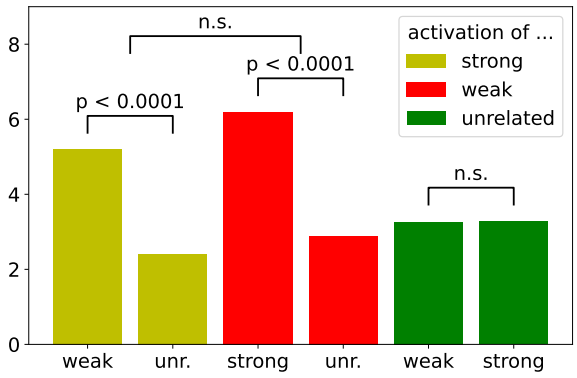

Figure 2: Scalar activation of adjectives after various prompts (OPT 125m). Individual bars show conditioning of the scalar terms.

ticipants were presented with isolated lexical items. To check whether LMs are sensitive to this, we repeat all of the above settings such that the LM is presented with the adjectives in isolation.

2.1 Scale selection

We use experimental materials from Lacina and Gotzner (2024), which contains constant sentence frames and focuses on one grammatical class (adjectives). To mitigate tokenization effects, we exclude 18 of 64 scales where any of the adjectives is split into more than one subword token.

3 Activation for language models

We use next token prediction models, which assign weights (logits) that indicate how well a token is activated by the context (Fig. 1). The softmax function transforms the logits into a probability distribution over the vocabulary, which is used for next word prediction. As an effect, tokens that are ranked among the top 10% of continuations receive low probabilities (see adjectives in Fig. 1). Compared to softmax probabilities, logits for individual tokens are relatively independent from each other. We calculate the activations of strong adjectives from both probabilities and logits, and use the paired sample t-test for the condition effect for the strong adjective activation. We find that the effect is significant for logits (p<0.0001) but not probabilities (p=0.21). This finding aligns with the research on the internal prediction construction process of LMs (Geva et al., 2022). In what follows, we use logit values as the activation measure.[2]

4 Results

We test OPT (Zhang et al., 2022) models of varying sizes from 125M to 6.7B parameters. All but one model demonstrate similar behaviour both with and without context (Tab. 1). Fig. 2 presents results for the smallest model: Activation of strong and weak adjectives is significantly higher after a related adjective; activation difference and activation of unrelated adjectives do not vary significantly.

The only exception is the second smallest (350M) model. Because we do not have insights into the training process of the models, we refrain from making claims about the reason for this unexpected behaviour.

5 Discussion

The presented setup allows to study the activation of vocabulary items beyond discrete token predictions. This allows to test whether linguistic concepts (e.g., scalar activations) are captured by the LM. As a next step, we will examine scalar activation in more complex contexts as in Sun et al. 2018 and Nizamani et al. 2024, and track the development of activations at several points in the sentence. We will also test whether linguistic features of the scales (e.g., boundedness) correlate with the magnitude of the activation effect for LMs, and whether the difference between the models is reflected in fine-tuning results.

[2]The absolute logit value depends not just on the vocabulary item, but also other factors such as sentence length. We subtract the mean of the logits over the vocabulary for presentational reasons. This does not affect the significance of the described effects.

References

Alex De Carvalho, Anne C Reboul, Van der Henst, Anne Cheylus, Tatjana Nazir, et al. 2016. Scalar implicatures: The psychological reality of scales. *Frontiers in psychology*, 7:203305.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicole Gotzner and Jacopo Romoli. 2022. Meaning and alternatives. *Annual Review of Linguistics*, 8(Volume 8, 2022):213–234.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, 9:1659.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Radim Lacina and Nicole Gotzner. 2024. Exploring scalar diversity through priming: A lexical decision study with adjectives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795, Torino, Italia. ELRA and ICCL.

Eszter Ronai and Ming Xiang. 2022. Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung*, volume 26, pages 716–733.

Eszter Ronai and Ming Xiang. 2023. Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, 2:229–240.

Chao Sun, Ye Tian, and Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology*, 9:2092.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.