Pre-trained Language Models' Interpretation of Evaluativity Implicature: Evidence from Gradable Adjectives Usage in Context

Yan Cong
yancong222@gmail.com

Abstract

By saying *Maria is tall*, a human speaker typically implies that Maria is *evaluatively* tall from the speaker's perspective. However, by using a different construction *Maria is taller than Sophie*, we cannot infer from Maria and Sophie's relative heights that Maria is evaluatively tall because it is possible for Maria to be taller than Sophie in a context in which they both count as short. Can pre-trained language models (LMs) "understand" evaulativity (EVAL) inference? To what extent can they discern the EVAL salience of different constructions in a conversation? Will it help LMs' implicitness performance if we give LMs a persona such as chill, social, and pragmatically skilled? Our study provides an approach to probing LMs' interpretation of EVAL inference by incorporating insights from experimental pragmatics and sociolinguistics. We find that with the appropriate prompt, LMs can succeed in *some* pragmatic level language understanding tasks. Our study suggests that socio-pragmatics methodology can shed light on the challenging questions in NLP.

1 Introduction

This paper concerns pre-trained Language Models' (LMs) interpretation of context-specific implicit elements on the pragmatic level of language understanding. Probing LMs' competence in implicitness is challenging due to the lack of surface representation. In this paper, we attempt to tease apart exactly what LMs "know" about pragmatics through a case study of gradable adjectives such as *tall*. We draw insights from experimental pragmatics and sociolinguistics, and implement them in probing two types of transformer LMs: the traditional auto-regressive GPT-3 (Brown et al., 2020) and the encoder-decoder model Macaw (Tafjord and Clark, 2021). Our findings show that the extent to which LMs are sensitive to implicitness depends on adjective properties (class, polarity, construction), prompt setting (the speaker is predefined as chill or nerdy), and transformers' architecture (decoder-transformer such as GPT-3, encoder-decoder transformer like Macaw).

By uttering a positive construction (henceforth POS) *Alex is tall*, conversational participants simultaneously extract two kinds of meaning: its descriptive literal meaning about the state of the world - Alex's height is above a particular threshold (Cresswell, 1976; von Stechow, 1984; Bierwisch, 1989); its socio-indexical meaning which implicitly reveals about the speakers themselves - Alex is tall from the speaker's perspective, namely the speaker *implies* that Alex is *evaluatively* tall (Bierwisch, 1989; Rett, 2008a,b). By contrast, when uttering an equative construction (henceforth EQ) like *Alex is as tall as Arthur*, or a comparative construction (henceforth COMP) such as *Alex is taller than Arthur*, there is no such salient *evaluative* reading because it's likely that Alex is as tall as or taller than Arthur in a context where (the speaker thinks) they are both short. A construction is evaluative if and only if it contextually entails its POS counterpart (Bierwisch, 1989; Brasoveanu and Rett, 2018). This is called the Bierwisch Test: by using *Alex is tall* (POS), the speaker implies that Alex is *evaluatively* tall; while by using *Alex is as tall as Arthur* (EQ) or *Alex is taller than Arthur* (COMP), the speaker is not implying that Alex is *evaluatively* tall - hence the linguistic generalization: using POS gives rise to evaluativity (henceforth EVAL) implicatures, whereas using EQ or COMP does not. We make all code and test data available for additional testing [1]

EVAL is a central member of the class of context-sensitive phenomena. It arises as a pragmatic inference - a conversational implicature (Rett, 2015, 2019; Bumford and Rett, 2020). Our paper proposes LMs examination schemes through a case study on EVAL implicature. Our study is built up

[1] https://github.com/yancong222/Unimplicit

Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language, pages 1 - 7
July 15, 2022 ©2022 Association for Computational Linguistics

on Brasoveanu and Rett (2018), which adopts the Bierwisch Test to test for the the presence of EVAL inference in different gradable adjectives. Their experimental pragmatics findings show a comprehensive picture about EVAL implicature (human judgment data, N=95): humans think that EVAL implicature is highly dependent on *context* which is shaped by the speaker's usage of adjective class (relative, e.g. *heavy*, absolute, e.g. *full*), adjective polarity (positive: *tall*, negative: its antonym *short*), and construction (POS, EQ, COMP). Their experiment result (Table2) showed that regarding adjective polarity, there is no clear difference in EVAL between positive and negative adjectives within either the relative or the absolute class. Regarding construction, POS is clearly the most evaluative. Regarding adjective class, the relative adjective class is less evaluative than the absolute in POS, but more evaluative than the absolute in EQ, and exhibit the same EVAL as the absolute in COMP. Our LMs investigation implemented Brasoveanu and Rett (2018)'s dataset to examine the extent to which LMs align with humans.

Throughout Brasoveanu and Rett (2018)'s dataset, only one template prompt was used. Thus, as a sanity check, we varied the prompt by adding two distinct personality illustration to the input. Another motivation of taking prompt design to be an independent variable is that an utterance's socio-indexing meaning and its speaker's personality traits are intertwined: it reveals about the speakers' demographic background and ideological orientation (Labov, 2006; Silverstein, 2003; Eckert, 2008; Podesva, 2011). We designed the prompt text based on *speaker persona*: a social construct shown to be central to social meaning across various domains of language (Eckert, 2008; Podesva, 2011). We argue that the construct of persona is relevant to LMs examination because: i) it's well-known and readily available for perceiving social identity in human interaction; ii) it's been shown to shape human language processing at different levels (Niedzielski, 1999; Strand, 1999; Casasanto, 2008; Choe et al., 2019); iii) personae tend to be indexed by a variety of (non-)linguistic signs, including a mere textual description of the persona at stake (D'onofrio, 2018), making them easy to invoke in LMs experiment set up.

Specifically, our incorporation of *persona* in the prompt design is inspired by Beltrama and Schwarz (2021). They find that to compute the standard of precision required to interpret numeral expressions, human comprehenders reason about the speaker's social identity, particularly about the persona they embody. An utterance produced by a nerdy speaker is associated with higher standard of precision, hence the tendency to interpreting the literal meaning but not necessarily the socio-indexing pragmatic implicit meaning, compared to the same utterance in the same context uttered by a chill speaker. Our experiments on LMs take two opposite sets of characters: a persona interpreting utterance with its literal meaning (Nerd), and a persona embodying laid-backness and pragmatic skillfulness (Chill). We framed these two persona in the prompt text, and examined if this could help LMs "understand" EVAL implicatures across various adjectives. We found that the answer depends on adjective properties and LMs' types.

A lot of attention has been paid to increase LMs' general transparency (Ettinger, 2020; Rogers et al., 2020), among which studies on LMs' interpretation of implicitness mostly focus on scalar implicature or presupposition (Schuster et al., 2020; Jeretic et al., 2020; Pandia et al., 2021). To our knowledge, no studies in this line have been done on gradable adjectives' EVAL implicature, although EVAL and gradability are classic topics in context sensitivity. This is probably because these phenomena are cognitively too subtle to spot, hence hard to quantify under a LMs framework.

Against this background, our goal is to examine the extent to which pre-trained LMs can "understand" implicit EVAL implicatures. We hypothesized that if pre-trained LMs are cognitively plausible, their performance should align with the human data in Brasoveanu and Rett (2018) and Beltrama and Schwarz (2021), namely: i) there should be no EVAL difference regarding adjective polarity, ii) LMs should predict POS constructions to be the most evaluative, iii) whether LMs consider the relative adjectives to be more or less evaluative than the absolute adjectives depends on construction type, iv) LMs should (at least) show a trend that the chill-persona prompt helps LMs' understanding of implicatures, relative to the nerdy-persona prompt.

2 Experiments

We designed our tests in the form of completion tasks, so as to test the pre-trained LMs in their most natural setting, without interference from fine-tuning. We presented all the tasks in a conversation format involving agent(s), meaning LMs are expected to interpret the utterance with some conversational level of language understanding. We focus on two distinct types of transformers (Table1): Macaw (Tafjord and Clark, 2021), which is more recent (built on top of T5 Raffel et al. (2020)), and GPT-3 (Brown et al., 2020). We used the 32 gradable antonym pairs (16 relative adjectives and 16 absolute adjectives) in Brasoveanu and Rett (2018), because it's already quantitatively justified by human judgments. Each antonym pair was syntactically framed in 3 distinct types of constructions: POS, EQ and COMP. This gave us 32 (adjectives) x 3 (constructions) = 96 strings of sequence.

Model	n_{params}	n_{layers}
Macaw-large (c.f. T5)	770M	24
GPT-3/InstructGPT	175B	96

Table 1: (pre-trained LMs) Model cards

Input representation We adapted Brasoveanu and Rett (2018)'s conversational prompt template involving multiple agents. LMs were presented with deductions a Police Chief (*agent*1) makes based on one-sentence utterance reports from his Detective (*agent*2) - *The Detective reported to the Police Chief: "Maria is as short as Sophie." What can the Chief conclude from this?*. LMs completed the prompt with a fixed max-length of sequence (*max_tokens*=100). We preset the penalty and the presence coefficients as 0.6, which were reasonable values if the aim is to just reduce repetitive samples (Brown et al., 2020). All the stimuli had the same format, the only strings that changed were the Detective's quoted report (underlined), which was replaced by different adjectival constructions.

In terms of prompt template, there were 3 variations: in addition to the Detective report, we adopted the Nerd versus Chill persona idea proposed and quantitatively justified by Beltrama and Schwarz (2021) (N=240). Their human data showed that Arthur, who is overwhelmingly seen as embodying social qualities indicative of nerd, is consistently associated with a geeky stereotype and tend to be insensitive to pragmatic cues, whereas Alex is ascribed attributes such as chill and a sociable personality, and he is pragmatically savvy. LMs were prompted with (1) Nerd persona: *Arthur is clever, smart, quiet, awkward, nerdy, shy and geeky. What does he mean by saying "Maria is tall"?* (2) Chill persona: *Alex is chill, laid-back, relaxed, easy, cool, friendly, and outgoing. What doe he imply by saying "Maria is tall"?*. All the adjectives used in the two persona prompts are from Beltrama and Schwarz (2021)'s collection of human responses to nerdy/chill stereotypes. All the stimuli had the same format, the only strings that changed were the speaker's (Alex or Arthur) quoted statement (underlined), which was replaced with various target adjectival constructions. The prompt examples are given in Table 3.

Measurement Inspired by the Bierwisch Test (Bierwisch, 1989), we hypothesized that a construction is evaluative if and only if it contextually entails its POS counterpart. We therefore used GPT-3 similarity embedding model *text-similarity-babbage-001* to embed document as a single vector (Brown et al., 2020). We deployed the model to both LMs' responses and the target utterance (the testing adjectival construction's POS counterpart). We then calculated the cosine distance of the two vectors. The similarity score is calculated only between LMs' own response and the target utterance (see Table 3 for examples).

Adopting Iter et al. (2018)'s semantic similarity metrics, where larger amounts of concept overlap between two text segments is interpreted as more similar, we computed the cosine similarity as a proxy to the measurement of the concept overlap between LMs' response and the target inference. We took that to be how much implicit meaning LMs can pick up in the conversation. For example, in the Detective setting with EQ in the Detective's quoted report "*Maria is as short as Sophie*", the target utterance is its POS counterpart *Maria is short*. Suppose LMs "understand" the EVAL implicature, LMs should draw a POS evaluative inference from EQ. This is reflected in the similarity: LMs' response is predicted to be similar to the target utterance if LMs makes the appropriate pragmatic inference.

3 Results and Discussion

With respect to **polarity** (Fig.1), the results align with human data. There is no statistically significant difference in EVAL between positive and negative adjectives within either the relative adjective class or the absolute adjective class. Regarding **constructions** (Fig.2), consistent with humans, the POS construction shows the highest similarity to the target inference: POS is the most evaluative across different LMs and adjective types. Regarding **LMs** in Fig.2, GPT-3 is more human-like than Macaw regarding construction sensitivity. GPT-3's output shows that using POS implies EVAL, using EQ is less likely to imply EVAL, and using COMP is the least likely to imply EVAL. By contrast, Macaw's output response to different constructions is not as stable: a lot of variance is found especially in Macaw - Nerdy interpreting EQ and COMP. Relative to GPT-3, Macaw is more sensitive to input instructions: with Chill personality, Macaw's "endorsement" of POS being evaluative gets significantly improved; whereas given nerdy personality (Macaw - Nerdy) or without any explicit identity, just interpreting Detective's report (Macaw - Detective), Macaw's sensitivity to constructions is not as salient. On the other hand, regardless of personality setup, GPT-3 showed similar patterns to different constructions.

With respect to **adjective class**: for POS (Fig.3 left), except for Macaw - Detective which considers relative adjectives to be *slightly* more evaluative than the absolute, LMs' output shows that the relative adjective class is less evaluative than the absolute adjective class. This effect is statistically significant for Macaw - Nerdy. Surprisingly, for both EQ (Fig.3 middle) and COMP (Fig.3 right), LMs still output representations suggesting that the relative adjectives are less evaluative than the absolute adjectives, especially for Macaw in which statistical significance was found. An exception was found in GPT - Detective in COMP, which judges relative as more evaluative than the absolute. GPT-3 did not seem to outperform Macaw, although t-test showed that GPT-3 did not *significantly* interpret absolute adjectives to be more evaluative than relative adjectives. In almost all of the cases, LMs indiscriminately "understood" absolute adjectives to be more evaluative than relative adjectives. Overall their interpretation of EVAL implicatures is not sensitive to construction.

Introducing socio-pragmatic frameworks in LMs evaluation loop, we adopted theory-driven hypothesis and cognitively justified datasets to analyze LMs' interpretation of EVAL implicature across adjective types. We found that LMs align with human data in that both suggest that polarity does not influence EVAL, and both considered POS to be the most evaluative across all adjective types, but deviant from linguistic theory and human cognition, most LMs' output suggests that the relative adjectives are *less* evaluative than the absolute across constructions. The persona setting helped *some* LMs "understand" implicitness. We provide an attempt to tackle challenging NLP questions using validated socio-pragmatic paradigms.

4 Limitation and Future studies

In this paper, we investigated the extent to which pre-trained transformer LMs (GPT-3 and Macaw) capture human inferences regarding the evaluativity of different adjectival constructions (POS, EQ, and COMP). We acknowledge that there are limitations, which we hope to address in future studies.

It might not be fine-grained enough to capture the extent to which LMs draw an evaluative inference using the cosine similarity measurement as a proxy. Specifically, our methodology design cannot account for the differences of the similarity scores between the target utterance (a) *Maria is tall* and (b) LM's response *Maria is taller than average*, and those between the target utterance (a) and an *irrelevant* distractor (b′) such as *The Detective is reporting on the height of Maria*, given the prompt *The Detective reported to the Police Chief: "Maria is tall." What can the Chief conclude from this?*. (b′) is entailed by the prompt, although it's not similar to (a) in a vector space. (b) is contextually entailed by the prompt but it's close to (a) in a vector space. This may distract LMs away from the target inference. For future study, we consider using a Natural Language Inference (NLI) model to more directly test the contextual entailment relationship between LMs' response and the target inference.

Typically, LMs are probed by looking at free form continuation or at the probability assigned to different output continuations under the LMs. In this paper, we probed LMs using a question-answering format and measured LMs' performance with similarity scores between LMs' answer and a target inference. But, even with the non-standard methodology, we still found some evidence that LMs do capture human inferences. Our study shows that it's worthwhile to adopt existing sets of human behavioral data in evaluating LMs' weakness and strength. Manipulating the social persona of the "speaker" in a LMs setting is novel yet at preliminary stage. For future work, we hope to expand our current dataset and better our motivation of the sociolinguistic persona design.

5 Appendices

	RELATIVE *tall*		ABSOLUTE *full*	
	POSITIVE *tall*	NEGATIVE *short*	POSITIVE *full*	NEGATIVE *empty*
POS	less evaluative *is tall*		more evaluative *is full*	
EQ	more evaluative *is as tall as*		less evaluative *is as full as*	
COMP	exhibit the same EVAL *is taller than* vs. *is more full than*			

Table 2: Distribution of EVAL implicature - human data: POS is the most evaluative; no difference in polarity (i.e., *positive* and its *negative* antonym) (Brasoveanu and Rett, 2018).

Prompts	GPT-3 response	Macaw response
The Detective reported to the Police Chief: "Maria is tall." What can the Chief conclude from this? [DETECTIVE]	The Police Chief can conclude that Maria is taller than average. The Detective is reporting on the height of Maria.	there's a difference. Maria is tall.
Arthur is clever, smart, quiet, awkward, nerdy, shy and geeky. What does he mean by saying "Maria is tall"? [NERD]	Maria is taller than Arthur. Maria is tall because she is taller than most people.	she's taller than he is. she is tall.
Alex is chill, laid-back, relaxed, easy, cool, friendly, and outgoing. What does he imply by saying "Maria is tall"? [CHILL]	Maria is tall implies that Maria is taller than most people. Maria is taller than Alex.	she is tall. Maria is tall.

Table 3: Example LMs completions with POS as prompt. EQ and COMP were situated in the same prompt frame.

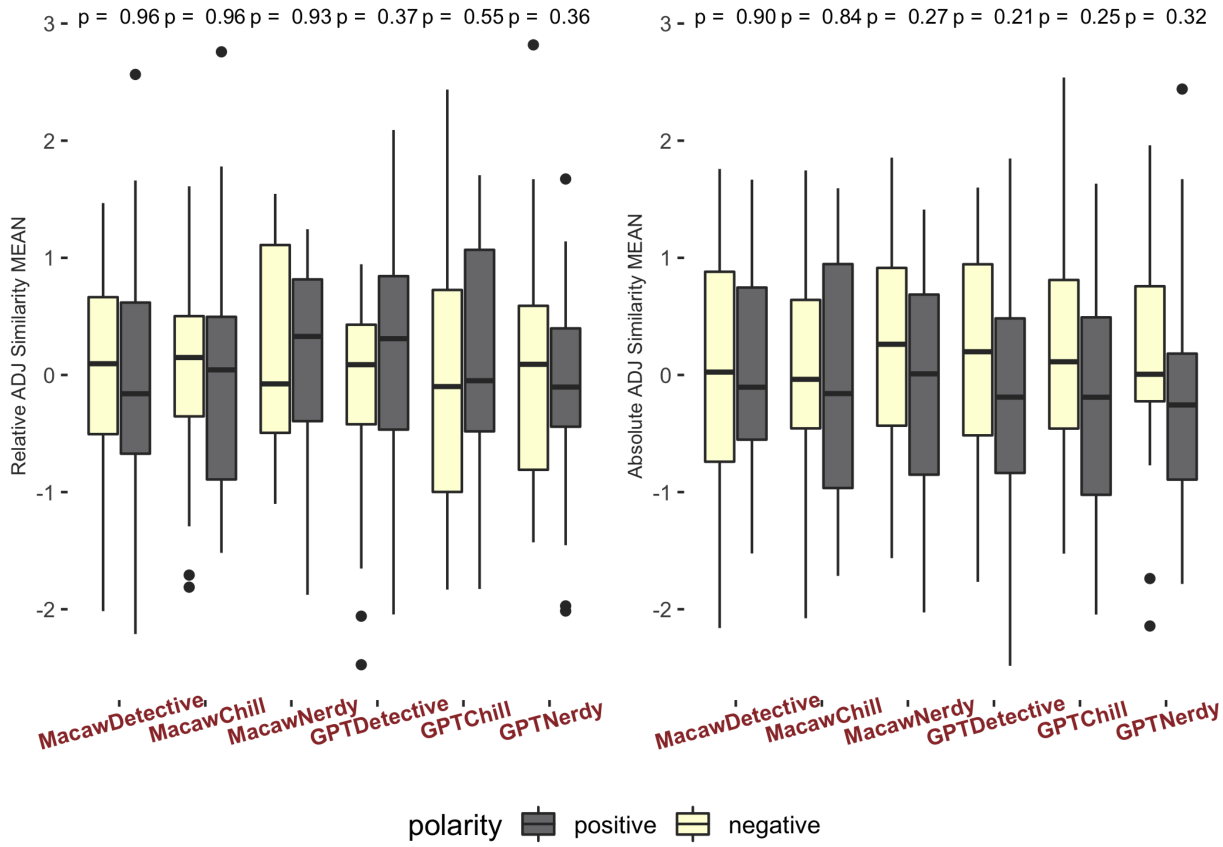

Figure 1: Polarity and LMs' understanding of EVAL. Panels split by adj_class. T-test *p*-values on the top.

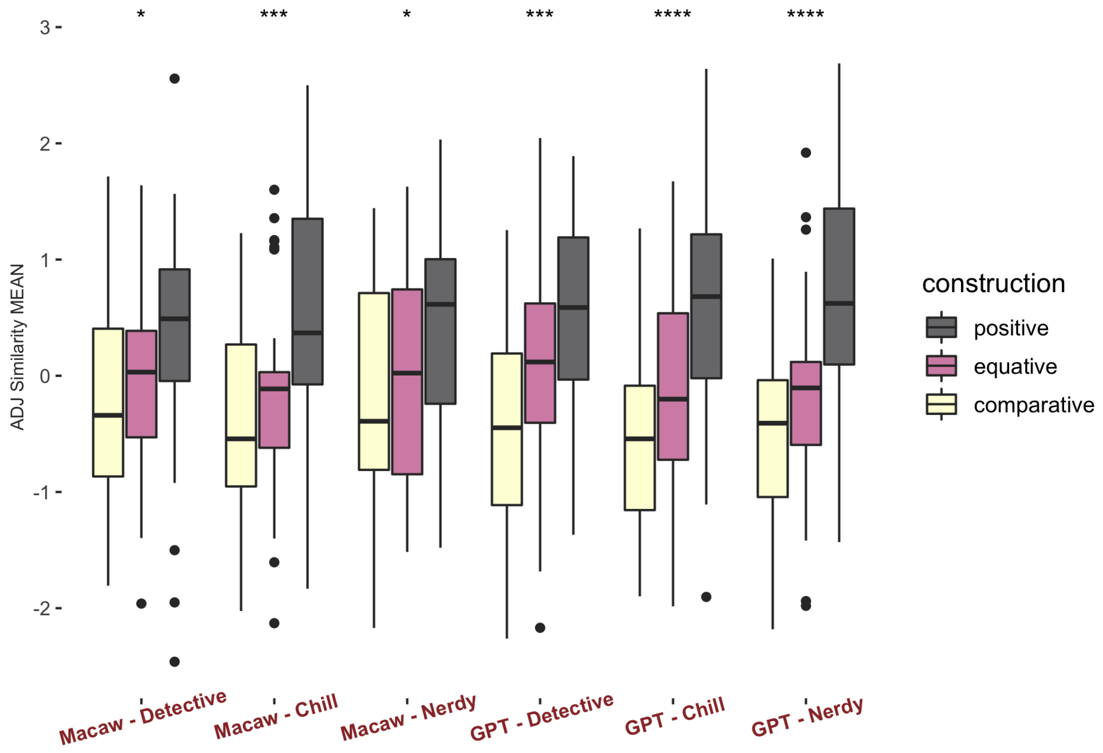

Figure 2: Construction and LMs' understanding of EVAL. Statistical significance (*.05*) on the top.

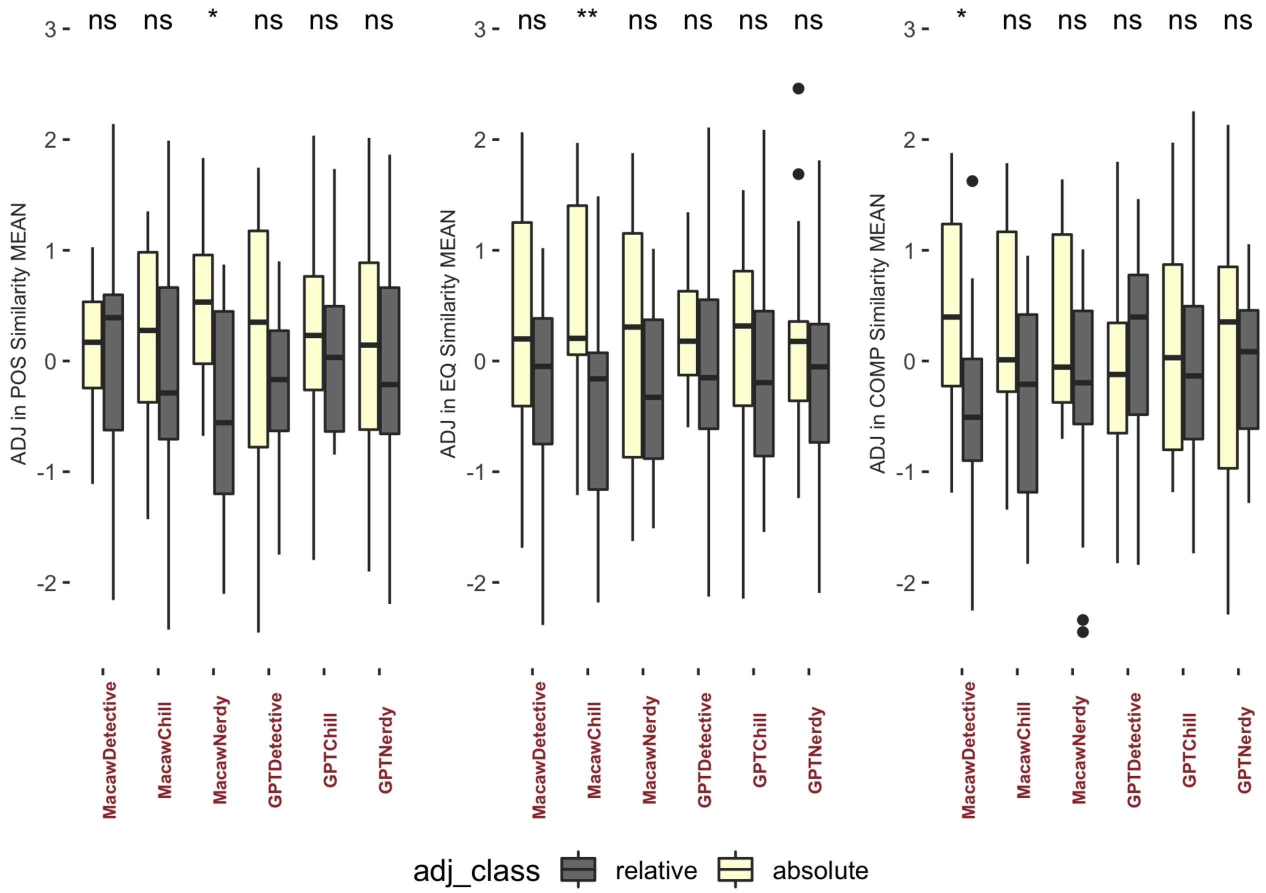

Figure 3: Adj_class and LMs' understanding of EVAL. Panels split by construction. Significance (*.05*) on the top.

References

Andrea Beltrama and Florian Schwarz. 2021. Imprecision, personae, and pragmatic reasoning. In Semantics and Linguistic Theory, volume 31, pages 122–144.

Manfred Bierwisch. 1989. The semantics of gradation. bierwisch, manfred & ewald lang (eds.), dimensional adjectives.

Adrian Brasoveanu and Jessica Rett. 2018. Evaluativity across adjective and construction types: An experimental study. Journal of Linguistics, 54(2):263–329.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Dylan Bumford and Jessica Rett. 2020. Rationalizing evaluativity. In Sinn und Bedeutung 25.

Laura Staum Casasanto. 2008. Does social information influence sentence processing? In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 30.

June Choe, Shayne Sloggett, Masaya Yoshida, and Annette D'onofrio. 2019. Personae in syntactic processing: Socially-specified agents bias expectations of verb transitivity. In Poster presented at the 32nd CUNY Conference on Human Sentence Processing.

M. Cresswell. 1976. The semantics of degree. In B.H. Partee, editor, Montague Grammar, 261-292. Academic Press.

Annette D'onofrio. 2018. Personae and phonetic detail in sociolinguistic signs. Language in Society, 47(4):513–539.

Penelope Eckert. 2008. Variation and the indexical field 1. Journal of sociolinguistics, 12(4):453–476.

Allyson Ettinger. 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.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 136–146, New Orleans, LA. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8690–8705, Online. Association for Computational Linguistics.

William Labov. 2006. The social stratification of English in New York city. Cambridge University Press.

Nancy Niedzielski. 1999. The effect of social information on the perception of sociolinguistic variables. Journal of language and social psychology, 18(1):62–85.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. arXiv preprint arXiv:2109.12951.

Robert J Podesva. 2011. Salience and the social meaning of declarative contours: Three case studies of gay professionals. Journal of english linguistics, 39(3):233–264.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

J. Rett. 2008a. Antonymy and evaluativity. In Proceedings of SALT XVII. CLC Publications.

J. Rett. 2008b. Degree Modification in Natural Language. Ph.D. thesis, Rutgers University.

Jessica Rett. 2015. The Semantics of Evaluativity. Oxford University Press.

Jessica Rett. 2019. Manner implicatures and how to spot them. International Review of Pragmatics, in press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8:842–866.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5387–5403, Online. Association for Computational Linguistics.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. Language & communication, 23(3-4):193–229.

Elizabeth A Strand. 1999. Uncovering the role of gender stereotypes in speech perception. Journal of language and social psychology, 18(1):86–100.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with Macaw. ArXiv, abs/2109.02593.

Arnim von Stechow. 1984. Comparing semantic theories of comparison. Journal of semantics, 3(3):1–77.