Shallowly accurate but deeply confused–how language models deal with antonyms[1]

Adèle HÉNOT-MORTIER — *Massachusetts Institute of Technology*

Abstract. Antonymic adjectives are subject to a variety of asymmetries regarding pragmatic inferences. The *Inference Towards the Antonym* (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018) in particular, consists in deriving the antonym of an adjective *A* when encountering its negation (*not* A). Within a given antonymic pair, this inference is supposed to apply to a greater extent to negated *positive* adjective, as opposed to negated *negative* adjectives. This is especially true when the latter is morphologically transparent. In this paper, we test if recent Large Language Models capture this contrast using different probing methods. We conclude that some but not all models exhibit a contrast between positive and negative adjectives regarding the target inference, although (i) the observed contrasts are not readily interpretable at the level of word processing (ii) part of it may be explained by frequency differences (iii) more general expectations about the models' behavior regarding antonymic adjectives (parsing, reversing effect of negation) are not met. This casts doubt on the ability of such models to abstractly encode the concept of antonymy.

Keywords: antonymic adjectives, polarity, pragmatic inferences, language models, surprisal

1. Background on adjective polarity

1.1. Semantics and pragmatics of antonymic adjectives

Antonymic adjectives, like (*tall*, *short*), (*nice*, *mean*), (*lucky*, *unlucky*) are roughly understood as semantic opposites. It has been observed that intuitively positive vs. negative adjectives pattern differently in several respects. First, only negative adjectives (abbreviated A^-) give rise to *Evaluativity Inferences* when used in equative and comparative constructions, as well as in questions Bierwisch (1989); Rett (2015). This is shown in (1) and (2).

(1) a. John is as tall_{A^+} as Paul. $\rightsquigarrow$ Both may be tall or short.
b. John is as short_{A^-} as Paul. $\rightsquigarrow$ Both are judged to be short by the speaker.

(2) a. How tall_{A^+} is John? $\rightsquigarrow$ John may be tall or short.
b. How short_{A^-} is John? $\rightsquigarrow$ John is judged to be short by the speaker.

Second, negative (rather than positive) adjectives may feature overt negative morphology (Horn, 1989). The examples in (3) below illustrate this point.

(3) a. in-competent; im-modest; un-lucky; dis-honest ...
b. *un-small; *im-messy; *un-poor; *dis-arrogant ...

Third, antonymic adjectives seem to differ in the inferences they lead to when placed under negation. Specifically, (4) shows that it appears easier to infer the antonym A^- of a negated positive adjective (abbreviated *not* A^+), than to infer the antonym A^+ of a negated negative adjective (*not* A^-) (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018).

[1]I would like to thank Forrest Davis and the students of Fall 2022 Special Seminar on Computational Linguistics for their comments on earlier versions of this project. I also would like to thank the audience and reviewers of Sinn und Bedeutung 28, as well as the attendees of the poster session and reviewers of the 29th Architectures and Mechanisms of Language Processing conference. All errors are my own.

(4) a. He is not tall$_{A^+}$. $\rightsquigarrow$ He is fairly short$_{A^-}$.
b. He is not short$_{A^-}$. $\not\rightsquigarrow$ He is fairly tall$_{A^+}$.

The inference in (4a) was dubbed *Inference Towards the Antonym* (henceforth ITA); it will be the focus of this paper. An account of the ITA, due to Krifka (2007), is based on the idea that any two antonyms A and A' are pure logical opposites of each other, which means that by default (*not* A) $\equiv$ A' and (*not* A') $\equiv$ A. This implies (*not* A) $\vDash$ A' and (*not* A') $\vDash$ A, i.e. the ITA is a (logical) primitive. It can however be *mitigated* if A and A' vary in complexity. More precisely, if *not* A appears more complex than A', then there are good reasons to think that the speaker wanted to convey a meaning different from A' when uttering *not* A, i.e. (*not* A) $\nvDash$ A'. This is summarized in (5), where CPLX refers to a measure of formal complexity.

(5) ITA Pragmatic Mitigation Condition (Krifka, 2007)
(*not* A) $\nvDash$ A', if CPLX(*not* A) $\gg$ CPLX(A') $\diamondsuit$

This allows to explain how a contrast in ITA *can* arise, but does not yet predict *in which direction* it arises. Building on the additional assumption due to Büring (2007); Büring (2007) that all negative adjectives involve either overt or covert negation, Krifka derives the two equations in (6). Small caps NOT refers to morphological (and potentially covert) negation.

(6) Negative Adjectives Complexity Hypothesis (Büring, 2007; Büring, 2007)
$\forall$ A$^-$. A$^-$= NOT-A$^+$, therefore:

$$\text{CPLX}(\text{A}^-) = \text{CPLX}(\text{NOT-A}^+) \quad \sim \text{CPLX}(\textit{not}\ \text{A}^+) \qquad (\spadesuit)$$
$$\text{CPLX}(\textit{not}\ \text{A}^-) = \text{CPLX}(\textit{not}\ \text{NOT-A}^+) \gg \text{CPLX}(\text{A}^+) \qquad (\clubsuit)$$

($\spadesuit$) states that *not* A$^+$ and A$^-$ have the same degree of complexity. No mitigation should therefore occur for that pair, and the ITA should arise. In other words, *not* A$^+$ is expected to entail A$^-$. ($\clubsuit$) on the other hand, states that *not* A$^-$ is significantly more complex than A$^+$. Pragmatic mitigation should therefore arise, leading *not* A$^-$ and A$^+$ to have different meanings.

1.2. Previous experimental investigation of the ITA

In the study conducted by Ruytenbeek et al. (2017), the inference pattern presented in (4) was assessed in English and French using minimal pairs like (7). In such pairs, the presupposition trigger *too* is expected to lead to the inference that *not* $A^{\pm}$ in the first sentence and A$^{\mp}$ in the second sentence have similar meanings. Since this inference is licensed from *not* A$^+$ to A$^-$, but not so much from *not* A$^-$ to A$^+$, (7a) is expected to be more felicitous than (7b).

(7) a. John is not tall$_{A^+}$. Paul is short$_{A^-}$ too. (*not* A$^+$) $\rightsquigarrow$ A$^-$
b. # John is not short$_{A^-}$. Paul is tall$_{A^+}$ too. (*not* A$^-$) $\not\rightsquigarrow$ A$^+$

In addition to testing the experimental validity of the basic contrast, and how it would correlate with independent measures of adjective polarity, Ruytenbeek et al. (2017) compared morphologically opaque pairs (e.g. *tall*/*short*) to morphologically transparent ones (e.g. *lucky*/*unlucky*). The goal was to investigate a refinement of the previous theory, based on the hypothesis

that morphologically transparent pairs should lead to a stronger ITA contrast than morphologically opaque pairs. This stems from the idea that the decomposition A^-= NOT-A^+ is more salient when a negative adjective is transparent as opposed to when it is not–which in turn means that (♣) should hold even more unambiguously for morphologically transparent pairs. Therefore, a stronger contrast (signaled with a double hashmark) is expected in pairs like (8) as opposed to pairs like (7) above.

(8) a. John is not lucky$_{A^+}$. Paul is unlucky$_{A^-}$ too. (*not* A^+) $\rightsquigarrow$ A^-
b. ## John is not unlucky$_{A^-}$. Paul is lucky$_{A^+}$ too. (*not* A^-) $\not\rightsquigarrow$ A^+

Building directly on Ruytenbeek et al.'s study on human participants,we propose to test if some recent Large Language Models (henceforth LLMs) verify the two hypotheses laid out in (9). This is to our knowledge the first study of the ITA in the context of LLMs (though see Aina et al., 2019 for a study on negated antonymic adjectives on earlier models, and Cong, 2022 for a study on evaluativity and LLMs).

(9) **H1**: it should be easier to infer A^- from *not* A^+ than A^+ from *not* A^-.
H2: the contrast in ITA strength between (*not* A^+)/A^- and (*not* A^-)/A^+) is bigger with transparent ("T") pairs of adjectives as opposed to opaque ("O") ones.

2. Technical and methodological background

2.1. The Transformer architecture

Probabilistic models of language, being for the most part based on the Distributional Hypothesis (Harris, 1954), have been previously shown to display poor performances in rendering the meaning of antonymic adjectives, in particular w.r.t. their interaction with negation (Aina et al., 2019). Recent LLMs, which are based on the Transformer architecture (Vaswani et al., 2017) and in particular the concept of *attention*, supposedly allow for more complex contextual dependencies between words, and as such might better grasp the meaning of antonymic adjectives, and the functional behavior of negation. Such models are also based on a process called *tokenization*, which allows to break certain words into pieces (*tokens*). In the following we provide an overview of tokenization and multi-head self-attention.

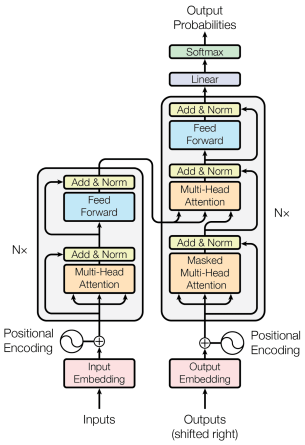
2.1.1. The tokenization process

Transformers operate on tokenized sentences, meaning, sentences whose words have been converted into one or several integers (*tokens*). Although it is not part of LLMs *per se*, tokenization remains crucial as it provides the models with interpretable inputs. The tokenization procedure relies on Byte-Pair Encoding (BPE), a process that creates tokens bottom-up from the set of characters (unigrams) appearing in the training corpus (initial workspace), by iteratively (i) merging the most frequent bigram, (ii) putting this bigram back as a unigram (with its corresponding frequency) in the workspace. The process stops once a specific vocabulary size has been reached. Since BPE is based on *n*-gram frequencies, it is expected to capture a certain number of morphological regularities. For instance, the existence of the negative morpheme *un-* in English probably makes the frequency of the corresponding bigram (*u*+*n*) comparatively high, making it likely to be categorized as a complete token. We will see in Section 4.2 that this kind of prediction is at least partly borne out for the adjectives involved in our dataset. If BPE is "productive' in the sense that any new word can be tokenized using its output vocabulary,

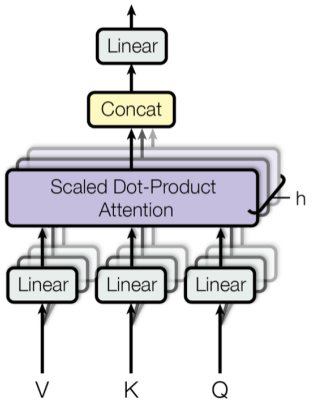
supplemented by the initial set of characters, and an extra "unknown" token for characters that did not belong to the initial set, this also entails that not all tokenizations will fully correspond to sensible morphological decompositions, either because some relevant morphemes are not identified, or because they are mistakenly identified in unexpected positions.

2.1.2. Multi-head self-attention

The main innovation of Transformers is the use of attention mechanisms, more specifically multi-head self-attention, as a core component of the network. Self-attention is a process that maps the representation of a given token t_j to an optimized mixture of the representations of the n surrounding tokens $\{t_i\}_{i\in[1;n]}$; the desideratum being that the weights of the mixture reflect how "relevant" those tokens are to t_j. *Multi-head* self-attention runs several such mechanisms ("heads") in parallel, allowing to capture different kinds of dependencies between tokens.



(a) The Transformer Encoder-Decoder architecture. Note that BERT only uses the encoder part, while GPT-2 only uses the decoder part.



(b) Detail of the multi-head self-attention architecture.

Figure 1: The Transformer architecture (taken from Vaswani et al., 2017).

Each head works as follows. First, the tokens $\{t_i\}_{i\in[1;n]}$ of the sentence are transformed ("embedded") into vectors $\{v_i\}_{i\in[1;n]}$ of dimension d_e. The goal of the self-attention head is then to map $\{v_i\}_{i\in[1;n]}$ to another set of vectors $\{y_i\}_{i\in[1;n]}$ containing more contextual information about each other. This mapping relies on three main sets of parameters, packaged into three matrices whose weights are subject to optimization and vary for each different self-attention head: the Query matrix $\mathscr{Q}_{(d_k\times d_e)}$, the Key matrix $\mathscr{K}_{(d_k\times d_e)}$ and the Value matrix $\mathscr{V}_{(d_v\times d_e)}$. Focusing on one input token-vector v_j and its target contextual representation y_j, v_j is first transformed into a d_k-dimensional query vector q_j using $\mathscr{Q}$. n d_k-dimensional keys are obtained by multiplying each of the $\{v_i\}_{i\in[1;n]}$ by $\mathscr{K}$. A dot product is then performed between the query q_j and each of the keys to obtain a list of scalar numbers that are subsequently normalized to yield the weights $\{w_{ji}\}_{i\in[1;n]}$. Finally, n d_v-dimensional "values" are obtained by multiplying each of the $\{v_i\}_{i\in[1;n]}$ by $\mathscr{V}$, and those values get linearly combined together using the weights $\{w_{ji}\}_{i\in[1;n]}$. This mixture of values is itself a d_v-dimensional vector, namely y_j, the target contextual representation of t_j. This whole series of operations can be performed for all $j \in [1;n]$, and for

each attention head $\{h_l\}_{l\in[1;m]}$, which leads to the more compact set of equations below. Note that the matrices $\mathscr{Q}$, $\mathscr{K}$, $\mathscr{V}$ now covary with the attention head h_l to model different kinds of contextual dependencies, and that the outputs of the m heads are combined and weighted by a matrix $\mathbf{W}$. Moreover, as Figure 1 shows, several (N) multi-head self-attention modules actually get *stacked* in the global architecture.

$$\mathbf{E} = \left[\left[\begin{bmatrix} \\ v_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ v_N \\ \\ \end{bmatrix}\right]\right] \qquad \begin{array}{rcl} \mathbf{Q}_l & = & (\mathscr{Q}_l\mathbf{E})^T \;:\; N\times d_k \\ \mathbf{K}_l & = & \mathscr{K}_l\mathbf{E} \;:\; d_k\times N \\ \mathbf{V}_l & = & (\mathscr{V}_l\mathbf{E})^T \;:\; N\times d_v \end{array}$$

$$\forall l\in[1;m].\; h_l = \text{softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l}{\sqrt{d_k}}\right)\mathbf{V}_l$$

$$\textsc{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [h_1 \dots\; h_m]\mathbf{W}$$

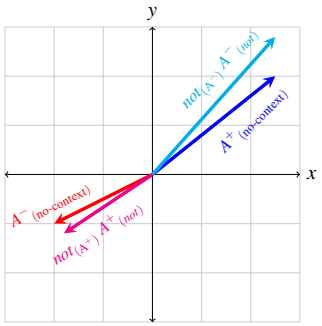


Figure 2: Idealized representation of the word-vectors of an antonymic pair and of their respective negations.

In practice, the output of vanilla LLMs is generative, which means that LLMs predict tokens given a certain context, by assigning them probabilities. Left-to-right models (e.g. the GPT family, cf. Radford et al., 2018; Radford et al., 2019; Brown et al., 2020) compute token representations and probabilities in a left-to-right fashion, while bidirectional models (e.g. the BERT family, cf. Devlin et al., 2018; Liu et al., 2019) do so using both left and right contexts.

2.2. The challenge of negated antonymic adjectives

Antonymic adjectives, and more so negated antonymic adjectives, pose a double problem to statistical models of language. The first problem is that of *grounding* (cf. Bender and Koller, 2020 a.o.). Since LLMs are simply trained to predict tokens, it is notoriously hard for them to capture intuitions about properties of the physical world, such as weight or size (though see Grand et al., 2022 for a discussion on the achievement of earlier models on nominals). This of course is problematic for adjectives, since many of them are highly context-dependent, and elements of an antonymic pair will often appear in similar distributional environments (Charles and Miller, 1989; Justeson and Katz, 1991). The second issue comes from the effect of negation. In formal semantics, negation is typically seen as a function that takes a proposition or predicate as argument, and return its opposite (proposition with opposite truth conditions, or complement set). LLMs however, treat any word as a vector, and therefore there is no way to properly "apply" the representation of negation to that of its argument in order to "reverse" it. One way for negation to alter its argument is in fact attention: as outlined in the previous section, LLMs derive contextual representations of words within a given sentence, so, ideally, we might expect negation to modify the representation of the adjective (and vice versa, in bidirectional architectures) in such a way that the representation of the negated adjective (typically seen as the mean of the representations of its constitutive tokens) becomes more or less close to the representation of its antonym, depending on polarity. This is schematized in Figure 2, where indices represent the context (assumed bidirectional) used to derive the representation of each token. Note however that this idealization puts a very high pressure on the contextual

aspect of representations: if $\overrightarrow{not\ _{(\mathrm{A}^{\pm})}\mathrm{A}^{\pm}_{(not)}} = 1/2\left(\overrightarrow{not_{(\mathrm{A}^{\pm})}} + \overrightarrow{\mathrm{A}^{\pm}_{(not)}}\right) \simeq \overrightarrow{\mathrm{A}^{\mp}}$ [2] then the contextual representation of *not* given any adjective of the antonymic pair is $\overrightarrow{not_{(\mathrm{A}^{\pm})}} \simeq 2\overrightarrow{\mathrm{A}^{\mp}} - \overrightarrow{\mathrm{A}^{\pm}_{(not)}}$ and the difference between the two contextual representations of *not* becomes proportional to the difference of the context-free representations of the two antonyms, which arguably is non-negligible: $\overrightarrow{not_{(\mathrm{A}^{+})}} - \overrightarrow{not_{(\mathrm{A}^{-})}} \simeq 3\left(\overrightarrow{\mathrm{A}^{-}} - \overrightarrow{\mathrm{A}^{+}}\right)$. We will see in Section 3.4 that this kind of constraint on contextual representations is not satisfied by the LLMs under study. The next two sections introduce two ways of probing the capacity of LLMs to successfully encode the semantics of adjectives.

2.3. Evaluating the linguistic performance of LLMs with surprisal

In human studies, the negative log-probability (surprisal) of a given word in a given context was shown to correlate with general processing effort (Hale, 2001; Levy, 2008). By extension, surprisal was taken as a reasonable proxy for syntactic acceptability when investigating the "linguistic" behavior of statistical models of language (Wilcox et al., 2018; Futrell et al., 2019; Wilcox et al., 2023) w.r.t. a variety of phenomena, among which filler-gap dependencies and island effects. The assumptions of this line of work are summarized in the equations in (10), where t_i denotes a token and $C(t_i)$ its context. For left-to-right models, C will denote the left context only, while for bidirectional models, C will denote both the left and right context. We will use the same kind of methodology in this paper except that we will assume that the measure of ACCEPTABILITY defined in (10) can, in the sentences at stake, reflect pragmatic acceptability.

(10)
$$\textsc{Surprisal}(t_i, C(t_i)) = -\log\left(\mathbb{P}(t_i|C(t_i))\right)$$
$$\textsc{Acceptability}(t_i, C(t_i)) \simeq -\textsc{Surprisal}(t_i, C(t_i))$$
$$\textsc{Acceptability}(t_1 \ldots t_N, C) \simeq -\sum_{i=1}^{N} \textsc{Surprisal}(t_i, C(t_i))$$

2.4. Evaluating LLMs on logical inferences

It is also possible to evaluate certain LLMs on logical inferences without appealing to measures of surprisal. In that case, the models are fine-tuned to perform at specific kind of classification task called *Natural Language Inference* (NLI). Fine-tuning consists in keeping most of the parameters of the model untouched, while adding (and training) an extra final layer suited for the particular task at stake. For NLI, the task typically consists in deciding if two sentences are in a relation of logical entailment, contradiction, or logically independent, by outputing a probability. Although it appears more direct than a surprisal-based assessment, this kind of task relies on the capacity of LLMs to transfer a "knowledge" acquired on the general instances of entailment encountered during training, to the particular case of the ITA.[3]

3. Experiments

3.1. Setup

The code used for the experiments is available here. First, a dataset comprising 107 pairs of English antonymic adjectives was manually created. There was some degree of redundancy in the adjectives used across pairs, due to synonymy. For instance, the positive adjective *kind* was

[2] Given our hypothesis about the ITA, this last equality holds more for *not* A^{+} than for *not* A^{-}. This is illustrated in Figure 2: the vector of A^{-} is slightly closer to that of *not* A^{+} than the vector of A^{+} is to that of *not* A^{-}.

[3] This holds even more that one of the most popular NLI dataset used to fine-tune models, MNLI (Williams et al., 2018), contains very few instances of pure pragmatic inferences, as observed by Jeretic et al. (2020).

paired to the opaque negative adjective *mean*, but also, to the transparent negative adjective *unkind*. The dataset contained a total of 48 transparent ("T") pairs, and 59 opaque ("O") pairs. The experiments involved three main tasks. Task 1 focuses on surprisal measures to assess differences in ITA strength. Task 2 probes NLI models to directly measure ITA strength *via* entailment probabilities. Task 3 compares the contextual vector representations assigned by LLMs to antonyms and their respective negations to determine if contrasts in ITA strength translate into model-internal topological regularities. In Tasks 1 and 3, four models (all in their "Large" version from Huggingface) were tested: GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019). As mentioned earlier, GPT-2 is purely left-to-right; while BERT and RoBERTa are bidirectional. Lastly, XLNet features a left-to-right architecture but its objective function allows it to incorporate some bidirectional dependencies during training, which, arguably, makes it combine the best of both worlds.[4] In the second task, two models fine-tuned on the MNLI dataset (Williams et al., 2018) were assessed: RoBERTa (supposedly better than BERT) and DeBERTa (supposedly better than RoBERTa, He et al., 2020). Both were in their "Large" version.

3.2. Task 1: comparing measures of surprisals in minimal pairs

This task aimed at testing surprisal contrasts at the word- and sentence-level, on minimal pairs following the template in (11), inspired by Ruytenbeek et al.'s original stimuli.[5] All pairs of sentences were counterbalanced for gender (by swapping the pronouns) and "filled" with the 107 possible (A^+, A^-) antonymic pairs. For each minimal pair, we collected differences in sentence-level and word-level[6] surprisals using the Python `minicons` library (Misra, 2022).

(11) "Anteposed *too*" template
  a. He is not A^+. She too is A^-.
  b. # He is not A^-. She too is A^+.

(12) spells out how the hypotheses introduced in (9) translate in terms of total sentence surprisal

[4]Bidirectional models are expected to be overall better at modeling natural language, since not all kinds of dependencies are purely left-to-right. Those models however, are trained on a masked language modeling objective, which consists in replacing input tokens one at a time by a dummy token `MASK`, and learning to predict the original token using its bidirectional context. This causes this family of models to get worse at fine-tuning (which does not involve any artificial `MASK` token); and, also, this makes such models unable to capture joint probabilities in their prediction of the masked tokens, due to them being predicted only on the basis of their respective contexts.

[5]In addition to the template in (11), we tested a template in which *too* appeared after the second adjective, and a template without *too* but with the predicate *mean* coordinating the two propositions.

(i) "Postposed *too*" template
  a. He is not A^+, and she is A^- too.
  b. # He is not A^-, and she is A^+ too.

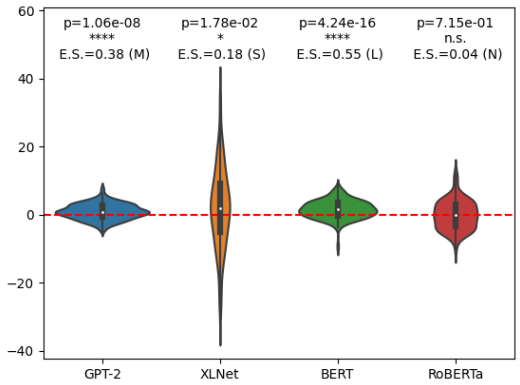
(ii) "Meta" template
  a. He is not A^+ means that he is A^-.
  b. # He is not A^- means that he is A^+.

The first template is closer to Ruytenbeek et al.'s original stimuli but is doing less justice to the left-to-right models (for which processing the presupposition trigger *too* before the second adjective might be crucial). The second template was used to neutralize the role of presuppositions in the semantic judgment, as it is yet unclear whether LLMs reliably "compute" presuppositions in the first place (Jeretic et al., 2020). Results for those templates can be generated using the project notebook, and do not fundamentally differ from those presented here.

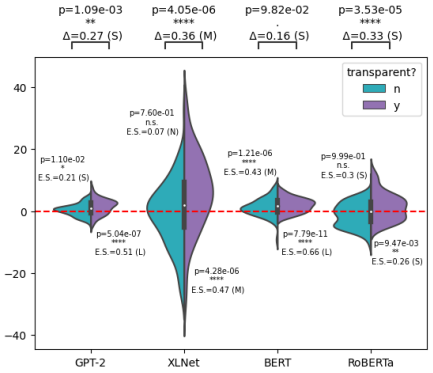
[6]If a word was segmented into several tokens, its surprisal was computed by simply summing the surprisals of its constituent tokens.

contrasts for the template in (11).

(12) **H1:** SURPRISAL(11b) - SURPRISAL(11a) > 0
H2: SURPRISAL(11b)$_{A\in T}$ - SURPRISAL(11a)$_{A\in T}$ > SURPRISAL(11b)$_{A\in O}$ - SURPRISAL(11a)$_{A\in O}$



(a) Testing H1. p-values[7] computed using one-tailed, Holm-Bonferroni-corrected Wilcoxon tests, effect size E.S. = $\frac{|z|}{\sqrt{n}}$.



(b) Testing H2 (T- vs. O-group). Within-group p-values computed using one-tailed, HB-corrected Wilcoxon tests; between-group using HB-corrected Mann-Whitney U-tests, and Cliff's Δ as effect size.

Figure 3: Paired differences in total sentence surprisal between (11b) and (11a).

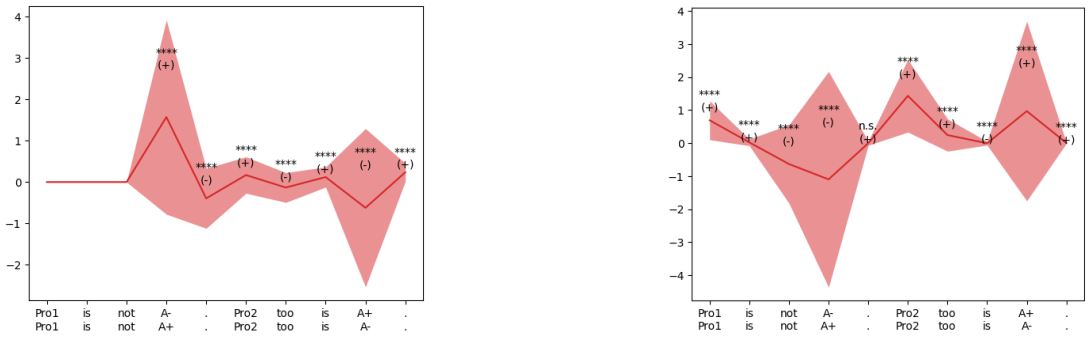
Figure 3a shows that all models but one (RoBERTa) exhibit a significant contrast in surprisal as a function of adjective polarity, with effect sizes varying from small to large.[8] This is in line with H1. Figure 3b shows that with GPT-2 and BERT, H1 is also individually verified by both the T- and O-group (with small to large effect sizes).[9] GPT-2 additionally appears to verify H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group, with a small effect size.[10] BERT only marginally verifies this prediction after corrections. This constitutes preliminary evidence that some LLMs capture the contrast in ITA strength *vis à vis* adjective polarity (H1) and its interaction with morphological transparency (H2). Remarkably, the two models that were supposedly more robust on general linguistic benchmarks, RoBERTa and XLNet, appear to perform less well than the basic models on this task.

Let us now focus on what the two best-performing models do at the word-level. From a language processing standpoint, we expect the positive contrasts in surprisal hypothesized at the sentence-level in (11) to be mainly driven by the occurrence of the second adjective. This adjective is expected to be relatively unsurprising when *negative* (due to the comparatively stronger ITA triggered by the negated *positive* adjective in the preceding sentence) and more surprising when *positive* (due to a weaker or absent ITA in the preceding sentence). This is summarized in (13) below, where A_2 refers to the second adjective in (11a)/(11b). This prediction however, might be influenced by whether the model under study is left-to-right or bidirectional. Indeed,

[7] p-value coding scheme: [.0001; −∞] ≡ ****; [.001; .0001] ≡ ***; [.01; .001] ≡ **; [.05; .01] ≡ *; [.1; .05] ≡ ·.
[8] This result is robust across templates for GPT-2 and BERT.
[9] Not 100% robust across templates: the O-group in the postposed *too* paradigm with GPT-2, and the T-group in the "meta" paradigm with BERT, failed to reach significance.
[10] Robust across all three templates–cf. footnote 3.2 for what the other templates were.

since in bidirectional models the conditional probability of each token is computed given its left and right context, the surprisal contrast expected on the second adjective might spread to other elements preceding it and "interacting" with it *via* attention, typically, the presupposition trigger *too*, the predicate *mean*, or the first adjective.

(13) **H1:** SURPRISAL(A_2, 11b) - SURPRISAL(A_2, 11a) > 0
H2: SURPRISAL(A_2, 11b)$_{A_2 \in T}$ - SURPRISAL(A_2, 11a)$_{A_2 \in T}$ >
SURPRISAL(A_2, 11b)$_{A_2 \in O}$ - SURPRISAL(A_2, 11a)$_{A_2 \in O}$



(a) GPT-2 (b) BERT

Figure 4: Paired word-by-word differences in surprisal between (11b) and (11a), *p*-values computed using Wilcoxon tests. The red line tracks the mean surprisal for each word and the red enveloppe tracks the standard deviation.

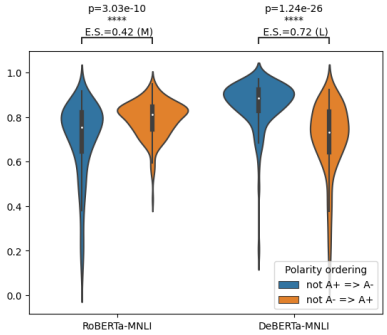
Figure 4a shows that GPT-2 treats A^- as significantly more surprising than A^+ in positive contexts (position 9, second adjective), but, even more so, in negative contexts (position 4, first adjective). The contrast in surprisal observed at the sentence-level for GPT-2 therefore seems to be driven by the first adjective, and not the second adjective, contrary to intuitions about the ITA, but perhaps consistent with a general avoidance for doubly negated (and therefore marked) structures, following Büring's Negative Adjectives Complexity Hypothesis. With BERT, Figure 4b shows that the pattern gets partly reversed: A^+ appears significantly more surprising than A^- in both positions, but, more remarkably perhaps, a surprisal contrast arises at the level of the subject of the second sentence (which remained the same in both sentences of a given minimal pair!). This is consistent with the idea that bidirectional models tend to "spread" surprisal contrasts to neighboring "relevant" tokens; however the reason why the subject pronoun should be relevant to the adjective polarity contrasts remains quite obscure. Other relevant candidates, such as the presupposition trigger *too*, show a significant, although comparatively smaller, surprisal contrast. In sum, a word-level assessment of surprisal contrasts for the best-performing models suggests that the global effect witnessed at the sentence-level was driven by elements of the sentence which intiuitively were not predicted to be triggering the linguistic contrast. This in turn suggests that LLMs may rely on more superficial cues (such as bare frequencies, and perhaps, a derived concept of markedness) to assign sentences probabilities. Before digging even further into the LLMs' contextual representations of antonymic adjectives, we explore in the next section another method of assessing the strength of the ITA in minimal pairs.

3.3. Task 2: comparing entailment probabilities between minimal pairs

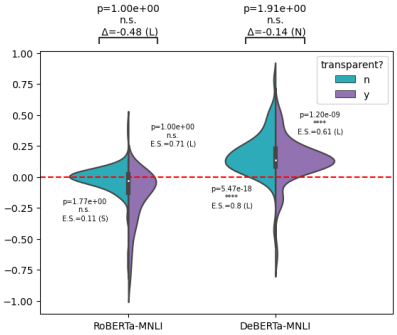
As outlined in Section 2.4 the contrast in (11) can be assessed using models fine-tuned to perform NLI. Those models are expected to associate the entailment patterns in (14) to a measure of probability reflecting how likely the relevant entailment is to hold for a particular pair of adjectives. (15) summarizes the specific predictions of (9) for this task, with $p^{\pm}_{A\in X}$ being the probability of entailment from *not* $A^{\pm}$ to $A^{\mp}$ when $A^{\pm}$ belongs to group X (T or O)

(14) a. He is not A^+ $\overset{p^+}{\Longrightarrow}$ He is A^-.
b. He is not A^- $\overset{p^-}{\Longrightarrow}$ He is A^+.

(15) **H1:** $p^+ - p^- > 0$
H2: $p^+_{A\in T} - p^-_{A\in T} > p^+_{A\in O} - p^-_{A\in O}$



(a) Testing H1. Probabilities of entailment for (14a) vs. (14b) on two LLMs fined-tuned for NLI. Same tests are in previous tasks.



(b) Testing H2. Paired differences in entailment probabilities between (14a) and (14b), T- vs. O-group. Same tests as in previous tasks.

Figure 5: Differences in entailment probabilities between (14a) and (14b).

Figure 5a shows that entailment scores are overall high for both models and both entailment schemes. Yet, only one of the two models (DeBERTa-MNLI) correctly predicts the inference in (14a) to be stronger than the one in (14b), in line with H1. The other model, RoBERTa-MNLI in fact predicts the opposite pattern. This negative result is consistent with the poor performance of the non-fine-tuned RoBERTa model in the previous task. Figure 5b shows that DeBERTa verifies H1 for the T- and O-groups individually (both with large effect sizes), but also that there is no significant difference in entailment strength between the two groups, which means that H2 fails to be supported. Overall, this inference task is not extremely explanatory, because it does not allow to determine if the models are drawing the desired inference for the "right" reasons. The next section is an attempt to better delineate what the basic models do under the hood, by analyzing the contextual representations of antonymic adjectives and their negations in the models' vector spaces.

3.4. Task 3: comparing vector representation of adjectives and their negations

Recall Figure 2, which illustrated what one should expect of two-dimensional, linguistically sensible contextual vector representations of antonymic adjectives and their negations. This Figure showed that A^+ and *not* A^- on the one hand, and A^- and *not* A^+ on the other hand, should cluster together and that, additionally, A^- and *not* A^+ should be closer to each other than A^+ and *not* A^-, due to the expected differences in ITA strength. The most common measure of semantic proximity used in word embeddings is cosine similarity, defined below,

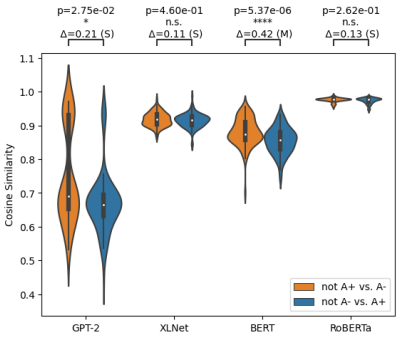
which corresponds to the measure of the angle between two vectors.[11] If H1 and H2 translate into the LLMs' contextualized vector space, we then expect the inequalities in (16) to hold.

(16) $\textsc{CosSim}(\vec{v_1},\vec{v_2}) = \frac{\vec{v_1}.\vec{v_2}}{||\vec{v_1}||\times||\vec{v_2}||} \in [-1;1]$

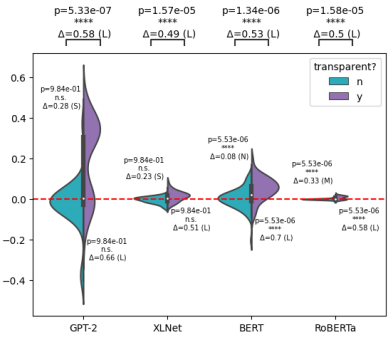
H1: $\textsc{CosSim}(\overrightarrow{\textit{not A}^+},\overrightarrow{\text{A}^-}) - \textsc{CosSim}(\overrightarrow{\textit{not A}^-},\overrightarrow{\text{A}^+}) > 0$

H2: $\textsc{CosSim}(\overrightarrow{\textit{not A}^+_{\text{T}}},\overrightarrow{\text{A}^-_{\text{T}}}) - \textsc{CosSim}(\overrightarrow{\textit{not A}^-_{\text{T}}},\overrightarrow{\text{A}^+_{\text{T}}}) > \textsc{CosSim}(\overrightarrow{\textit{not A}^+_{\text{O}}},\overrightarrow{\text{A}^-_{\text{O}}}) - \textsc{CosSim}(\overrightarrow{\textit{not A}^-_{\text{O}}},\overrightarrow{\text{A}^+_{\text{O}}})$

For each model, we constructed vectors for $\text{A}^{\pm}$ and *not* $\text{A}^{\pm}$, by averaging the representations of the second-to-last layer of the model obtained for each token.[12] Figure 6a shows that all models associate adjectives and the negation of their antonym to fairly high cosine similarities. GPT-2 and BERT moreover treat *not* A^+ and A^- as closer to each other than *not* A^- and A^+, with small to medium effect sizes, in line with H1 and the results of Task 1. Figure 6b additionally shows that BERT individually verifies H1 for both the T- and O-groups, as well as H2.



(a) Absolute cosine similarities for both adjective orderings. *p*-values computed using one-tailed, Holm-Bonferroni-corrected Wilcoxon tests, effect size E.S. = $\frac{|z|}{\sqrt{n}}$.



(b) Paired differences in cosine similarities, T- vs. O-group. Same tests and corrections as in previous tasks.

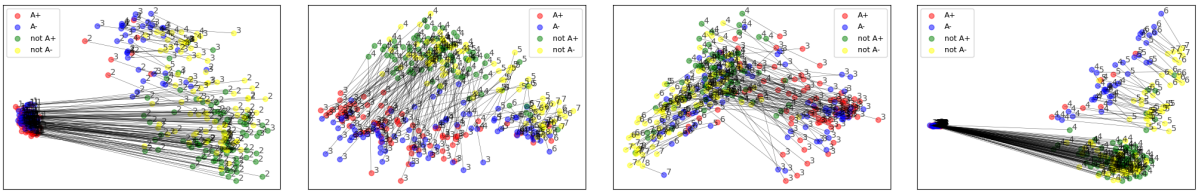
Figure 6: Differences in cosine similarities between $(\overrightarrow{\textit{not A}^+},\overrightarrow{\text{A}^-})$ and $(\overrightarrow{\textit{not A}^-},\overrightarrow{\text{A}^+})$.

These results are quite encouraging overall and suggest that the models which captured the desired surprisal contrasts in Task 1, encode antonymic adjectives and their negation in a somewhat sensible way, as well. This however, has to be nuanced with another fairly concerning aspect of the LLMs' contextual embeddings, visible in Figure 7 below, whereby bare antonyms (blue and red dots), and their negations (yellow and green dots) respectively end up clustering together in a 2D space where the dimensions that are retained are the ones that explain the most

[11] Two vectors pointing in the same direction will have a cosine similarity of 1, regardless of their respective lengths, while two vectors pointing in opposite directions will have a cosine similarity of -1. Orthogonal vectors have a cosine similarity of 0.

[12] Because some models tokenize words differently depending on whether they are preceded by a white space or not, we included an initial space before all the bare adjectives, to ensure they would be tokenized in the same way as they would be after negation. We also tried different vector extraction methods, in particular last-layer extraction (generally dispreferred due to the tendency of the last layer to encode information that is too task-specific) and summing of the last 4 layers (empirically better on certain benchmarks). Both methods led to comparable results as the one we retained in the main text, although the last-layer method led to slightly worse plots and *p*-values.

variance of the data. This clustering effect is evidently bigger than the one measured previously by comparing cosine similarities in higher dimensional spaces, and shows that the "reversing" effect of negation was not encoded by the models, thus replicating the negative result of Aina et al. (2019) for earlier models. Another concerning aspect of those 2-dimensional projections is the fact that the distributions of the vectors appear highly sensitive to the number of tokens they are derived from–this is particularly visible in the case of GPT-2 and RoBERTa for bare adjectives. This might also explain the bimodal aspect of the distribution of cosine similarities for GPT-2 in Figure 6, and calls for a more in-depth analysis of the LLMs' tokenization strategies.



(a) GPT-2 (b) XLNet (c) BERT (d) RoBERTa

Figure 7: Two-dimensional reductions (*via* Principal Component Analysis) of the contextualized representations of A^+, A^-, and their respective negations. The numbers indicate the total number of tokens (including start/end/separating tokens) each vector is derived from.

As an interim summary, it appears that some, but not all of the LLMs under study captured the effect of adjective polarity on the Inference Towards the Antonym, and did so, at the level of sentences (*via* surprisal measures) and at the level of contextualized word representations (*embeddings*). The measuring of word-level surprisals, as well as a broader analysis of the LLMs' contextual embeddings, however cast doubt on whether LLMs "draw" the target inference for the right reason. In the next section, we explore two potential confounding factors: adjective frequencies and possible biases caused by the tokenization procedure.

4. Analysis of confounding factors

4.1. Adjective frequency

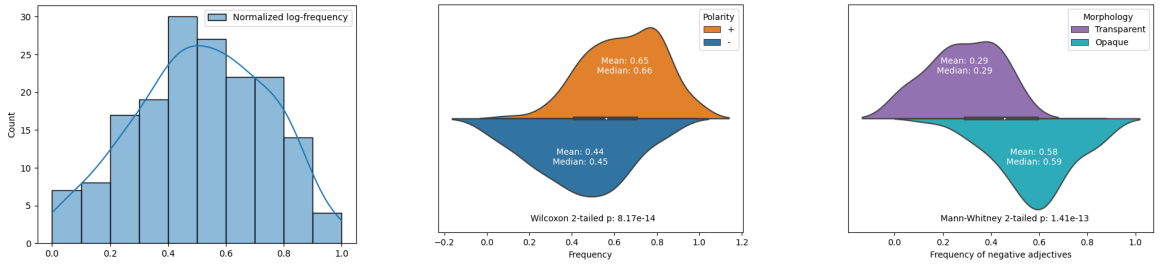
Since the training of Transformer models relies on statistical regularities, one might wonder if the effects observed are not just artifacts of frequency differences between positive vs. negative adjectives, and/or transparent vs. opaque adjectives. Can adjective frequencies explain the behavior of the LLMs under study w.r.t. Tasks 1 (surprisal) and 3 (inference)? To answer this question, we used a dataset from Kaggle[13] gathering the frequencies of the 1/3 million most frequent English words on the Web. This dataset was derived from the Google Web Trillion Word Corpus, distributed by the Linguistic Data Consortium (Brants, Thorsten and Franz, Alex, 2006). Even if the composition of this dataset might

Top 10 least frequent adjectives	Top 10 most frequent adjectives
ungraceful	just
uncommunicative	good
unambitious	well
unsocial	old
graceless	social
uncharitable	young
discourteous	popular
dextrous	fun
immodest	short
untruthful	bad

Table 1: 10 least and most frequent adjectives according to the Kaggle dataset.

[13] Dataset available at https://www.kaggle.com/datasets/rtatman/english-word-frequency.

differ from those of the datasets used to train the LLMs at stake,[14] we took it to be a sufficiently good approximation. This allowed us to extract the frequencies of all the adjectives from our dataset, which we further log-transformed and normalized.



(a) All adjectives (b) By polarity (c) By transparency

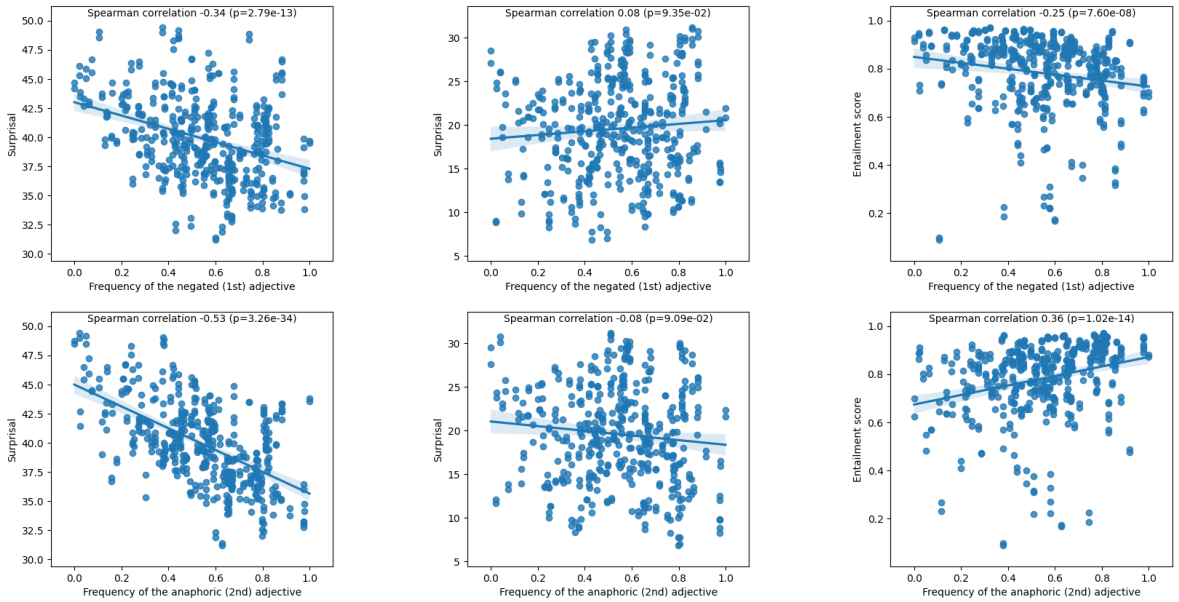
Figure 8: Distribution of the normalized log-frequencies of the adjectives from our dataset.

Table 1 and Figures 8a-8c illustrate the distribution of those normalized frequencies. Figure 8b shows that positive adjectives are overall more frequent than negative ones, within each pair (2-tailed paired Wilcoxon p=8.17e-14) as well as globally (2-tailed Mann-Whitney p=2.08e-11). This might be partly explained by the fact that more positive adjectives from the dataset have homonyms, and as such got their frequencies increased, as opposed to negative ones, which in almost half of the cases featured negative morphology specific to adjectival forms. *Just* and *well* in the top 10 most frequent positive adjectives in Table 1 are examples of such ambiguous positive adjectives. Figure 8c shows that within the class of negative adjectives, transparent ones appear less frequent on average than opaque ones (2-tailed Mann-Whitney p=1.41e-13). This again, might be partly explained by the potential for homonymy of O-adjectives.

Given these preliminary observations, we tried to assess the degree of correlation between total sentence surprisal measures (from Task 1) or entailment scores (from Task 2) on the one hand, and, on the other hand, the normalized log-frequencies of either the first (negated) adjective, or the second ("anaphoric") adjective in sentences like (11). To this end, we focused on the best-performing models for each Task. The intuitive expectation is that sentence surprisal measures should anti-correlate with the frequency of both adjectives, since surprisal convaries with the negative conditional probabilities of the tokens appearing in the sentence. Blocks (a) and (b) in Figure 9 show that this prediction is rather strongly verified for GPT-2, but not for BERT. This appears consistent with the word-by-word surprisal plots of Figure 4, which showed that the surprisal contrast with GPT-2 was driven by this model being overall more "surprised" at negative (i.e. less frequent) adjectives than positive (i.e. more frequent) ones, and that BERT weakly followed the opposite pattern. Regarding entailment scores, the prediction is less clear but we might expect more frequent adjectives in the conclusion to boost the probability of entailment. The lower plot of the (c) block in Figure 9 shows that this intuition is verified: when the adjective present in the conclusion becomes more frequent, the entailment score tends to increase, as well. It also seems that more frequent adjectives in the *premise* tend to make the entailment scores decrease (upper plot of the (c) block). This analysis suggests that GPT-2

[14] As an example, GPT-2 was trained on BookCorpus, which comprises 7,000 self-published independent books, and a curated Web corpus called WebText involving 8 million web pages. BERT was trained on BookCorpus and Wikipedia.

and DeBERTa may heavily rely on bare adjective frequencies to produce the desired contrasts in surprisal and entailment probabilities, respectively.



(a) GPT-2 (b) BERT (c) DeBERTa

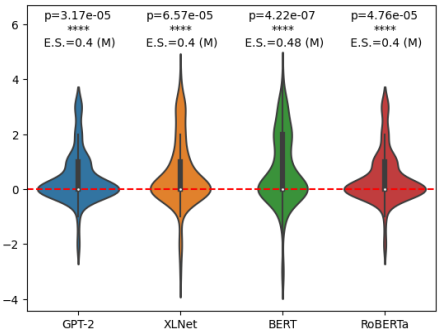
Figure 9: Correlation between adjective frequencies and total sentence surprisal scores (GPT-2, BERT) or entailment scores (DeBERTa).

4.2. Tokenization and morphology

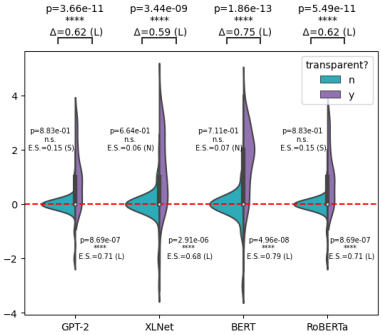
A last aspect of LLMs that may require further investigation is their tokenization procedure. As briefly outlined in Section 2.1, the input of Transformer models is a tokenized string, whose tokens may or may not coincide with actual morphemes. Tokenizers vary across models. Are the tokenized inputs formed out of our adjective dataset any close to morphologically-segmented data? Does the number of tokens of positive vs. negative adjectives reflect differences in formal complexity that can in turn influence surprisal or inference scores?

To answer these questions, we first computed, for each pair of adjectives, the differential number of tokens of A^- vs. A^+. Given that A^- is assumed to be overall more complex than A^+, the resulting differential number of tokens is expected to be positive. Figure 10 shows that this expectation is verified for all models, although the result seems to be driven by the transparent pairs only. This is not at all surprising given that tokenizers only have access to surface representations (strings) and as such cannot apply Büring's generalization to opaque pairs. We then tried to assess if differential numbers of tokens correlate with the surprisal contrasts measured on Task 1 for the two best-performing models (GPT-2, BERT); and the differential entailment scores measured on Task 3 for DeBERTa. The relevant scatter plots are shown in Figure 11 and suggest the existence of a weak positive correlation in the case of GPT-2, and a weak negative correlation in the case of DeBERTa. In other words, for GPT-2 the differential in complexity between A^+ and A^- tends to make the surprisal contrast between (11b) and (11a) bigger, which is somehow expected, while for DeBERTa, the differential in complexity between A^+ and A^- tends to make the contrast in entailment strength between (14a) and (14b) smaller, which is

unexpected. Differential numbers of tokens however, are perhaps not extremely informative if the parses generated by the tokenizers do not match the morphology of their input in the first place.

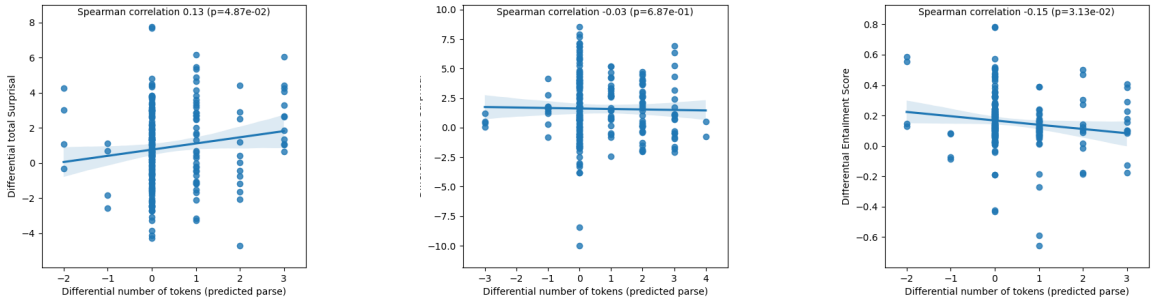


(a) All groups. Same tests and corrections as in previous tasks.



(b) T- vs. O-group. Same tests and corrections as in previous tasks.

Figure 10: Differential number of tokens between A^- and A^+.



(a) GPT-2 (b) BERT (c) DeBERTa

Figure 11: Correlation between differential number of tokens and differential surprisal scores (GPT-2, BERT) or entailment scores (DeBERTa).

For that reason, we assessed how accurate the tokenizers were in segmenting the adjectives from our dataset according to their actual morphological decomposition. In the general case, tokenizers managed to get the right parses between 42 and 48% of the time, but the accuracy significantly dropped when focusing on adjectives with plurimorphemic parses: GPT-2 and RoBERTa (which rely on the same tokenizer) only achieved a 4% accuracy, while BERT and XLNet respectively achieved 12 and 15%. Finally, we assessed how often tokenizations of morphologically transparent negative adjectives from our dataset involved a boundary between the negative morpheme and the base, since this decomposition is in theory the source of the complexity difference between positive and negative adjectives. We found that GPT-2 and RoBERTa only had a 21% accuracy in this particular task, while XLNet and BERT achieved a 60% accuracy. Those overall poor results imply that, even though some models exhibit the expected differences in complexity between A^+ and A^- in transparent pairs, and somehow rely on those differences to derive surprisal contrasts (in the case of GPT-2), they start out with representations that do not match linguistic theory.

5. Conclusion

We assessed various LLMs on their interpretation of antonymic adjectives and their respective negations, in particular, with regards to the Inference Towards the Antonym, which is expected to be stronger for negated *positive* adjectives as opposed to negated *negative* adjectives, and even more so for morphologically transparent pairs. Using measures of surprisal (Task 1), probabilities of entailment (Task 2) and vector similarities (Task 3), we found some evidence that two basic models (BERT and GPT-2), and one model fine-tuned for Natural Language Inference (DeBERTa) captured the predicted polarity contrast, and, in some cases, the magnifying effect of morphological transparency. More "advanced" models (on regular benchmarks) noticeably performed lass well on the tasks at stake. More targeted analyses however, showed that some reasonable expectations about the models' behavior were not met. In Task 1, even the LLMs which managed to give human-like "judgments" on minimal pairs, did not seem to focus on the right individual words to produce them and/or seemed to overly rely on bare adjective frequencies. In Task 2, we reported mixed results and, even for the best-performing model, observed correlations between entailment scores and frequencies of the target adjectives, as well as between differential numbers of token and differential entailment scores–which implies that the model might have relied on superficial cues to draw its conclusions. In Task 3, even when the LLMs' contextual representations appeared to capture ITA-related topological inequalities, the very same spaces were characterized by the stronger, very much unexpected topological regularity consisting in a clustering of bare antonyms on the one hand, and their negations, on the other hand. This clustering moreover seemed to depend on the number of tokens within each adjective.

In sum, some LLMs seem to be shallowly accurate in their treatment of antonymic adjectives, but also deeply "confused" about the sources of the relevant contrasts. More generally perhaps, this study questions how LLMs (and, in retrospect, humans!) can be sensitive to concepts such as markedness, and pragmatic competition. Should markedness be identified with formal complexity, and should differences in word frequencies be seen as the consequence of differences in markedness? Should the definition hold in the opposite direction? Or should markedness be seen as the result of an *interaction* between complexity and typicality? Finally, regarding competition and the nature of alternatives, it is worth nothing that the pragmatic framework we used makes the assumption that antonymic adjectives interact within a fixed *pair* but in practice, the negation of a given positive adjective might compete with more than one negative counterpart, and vice versa. This might make an account of the ITA more challenging, in that the *number* and relevance of potential competitors to a given negated adjective, in addition to the differential of complexity contributed by each competitor, might eventually play a role in the mitigation effect observed.

References

Aina, L., R. Bernardi, and R. Fernández (2019, June). Negated adjectives and antonyms in distributional semantics: not similar? *Italian Journal of Computational Linguistics 5*(1), 57–71.

Bender, E. M. and A. Koller (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 5185–5198. Association for Computational Linguistics.

Bierwisch, M. (1989). *The Semantics of Gradation*, pp. 71–261. Springer Berlin Heidelberg.

Brants, Thorsten and Franz, Alex (2006). Web 1t 5-gram version 1.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Büring, D. (2007, October). Cross-polar nomalies. *Semantics and Linguistic Theory 17*, 37.

Büring, D. (2007). More or less. In *Proceedings of the 43th Annual Meeting of the Chicago Linguistic Society*.

Charles, W. G. and G. A. Miller (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics 10*(3), 357–375.

Cong, Y. (2022, July). Pre-trained language models' interpretation of evaluativity implicature: Evidence from gradable adjectives usage in context. In V. Pyatkin, D. Fried, and T. Anthonio (Eds.), *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, Seattle, USA, pp. 1–7. Association for Computational Linguistics.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy (2019, June). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 32–42. Association for Computational Linguistics.

Gotzner, N., S. Solt, and A. Benz (2018, November). Adjectival scales and three types of implicature. *Semantics and Linguistic Theory 28*, 409.

Grand, G., I. A. Blank, F. Pereira, and E. Fedorenko (2022, April). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour 6*(7), 975–987.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

He, P., X. Liu, J. Gao, and W. Chen (2020). Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR abs/2006.03654*.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Jeretic, P., A. Warstadt, S. Bhooshan, and A. Williams (2020, July). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8690–8705. Association for Computational Linguistics.

Justeson, J. S. and S. M. Katz (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics 17*(1), 1–20.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. In U. Sauerland and

P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics*, pp. 163–177. London: Palgrave Macmillan UK.

Levy, R. (2008, March). Expectation-based syntactic comprehension. *Cognition 106*(3), 1126–1177.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving language understanding by generative pre-training.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

Rett, J. (2015). *The semantics of evaluativity*. Oxford: Oxford University Press.

Ruytenbeek, N., S. Verheyen, and B. Spector (2017, October). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: a journal of general linguistics 2*(1).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.

Wilcox, E., R. Levy, T. Morita, and R. Futrell (2018, November). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, pp. 211–221. Association for Computational Linguistics.

Wilcox, E. G., R. Futrell, and R. Levy (2023, 04). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44.

Williams, A., N. Nangia, and S. Bowman (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics.

Yang, Z., Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764.