Shallowly accurate yet deeply confused: how language models deal with antonyms

Adèle Hénot-Mortier (MIT)

September 7, 2023

Sinn und Bedeutung (SuB) 28

Introduction

Antonymic adjectives and the Distributional Hypothesis

- Antonymic adjectives, like (**tall**, **short**), (**nice**, **mean**), (**rich**, **poor**) are **semantic opposites of each other**.
- For that reason, they appear in **very similar distributional environments** (Charles & Miller, 1989; Justeson & Katz, 1991).
- 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).

Research question

- In this talk we are interested in how more recent "large" language models (**LLMs**) trained on large amounts of data deal with antonymic adjectives, in particular w.r.t. a certain kind of inference they trigger in negative contexts (**Inference Towards the Antonym**).
- Why? Recent LLMs are based on architectures which allow for complex **contextual** dependencies between words (or tokens). So they might be expected to better grasp the contextual meaning of antonymic adjectives, and the functional behavior of negation.
- **Studying such models will hopefully give us insights about how statistics are leveraged to approximate (or not!) human linguistic behavior.**

Roadmap

Introduction

Semantic background

Technical and methodological background

Task 1: "Behavioral" assessment based on surprisal measures

Task 2: "Inferential" assessment based on entailment probabilities

Task 3: "Internal" assessment based on vector similarities

Conclusion

Appendices

Semantic background

The basic contrast of interest

- Antonymic adjectives seem to differ in the inferences they lead to **when placed under negation**.

(1) a. He is not **tall**. $\rightsquigarrow$ He is **short**.

b. He is not **short**. $\not\rightsquigarrow$ He is **tall**.

- More specifically, it appears easier to infer the antonym $\mathbf{A}^-$ of a negated positive adjective (*not* $\mathbf{A}^+$), than to infer the antonym $\mathbf{A}^+$ of a negated negative adjective (*not* $\mathbf{A}^-$) (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018).
- The inference in (1a) has been called Inference Towards the Antonym (**ITA**), and will be the main focus of our study of antonyms through the lens of LLMs.

An account of the ITA: Krifka, 2007

The Inference Towards the Antonym (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018)

$$(\textit{not}\ \text{A}) \implies \text{A'}\ \text{where A and A' are antonyms.} \qquad (\heartsuit)$$

ITA Pragmatic Mitigation Condition (Krifka, 2007)

$$(\textit{not}\ \text{A}) \not\Rightarrow \text{A'}, \text{ if } \textsc{Cplx}(\textit{not}\ \text{A}) \gg \textsc{Cplx}(\text{A'}) \qquad (\diamondsuit)$$

Negative Adjectives Complexity Hypothesis (Büring, 2007a, 2007b)

$$\forall \mathbf{A^-}.\ \mathbf{A^-} = \textsc{Not-}\mathbf{A^+}, \text{ therefore:}$$

$$\textsc{Cplx}(\mathbf{A^-}) = \textsc{Cplx}(\textsc{Not-}\mathbf{A^+}) \sim \textsc{Cplx}(\textit{not}\ \mathbf{A^+}) \qquad (\spadesuit)$$

$$\textsc{Cplx}(\textit{not}\ \mathbf{A^-}) = \textsc{Cplx}(\textit{not}\ \textsc{Not-}\mathbf{A^+}) \gg \textsc{Cplx}(\mathbf{A^+}) \qquad (\clubsuit)$$

(1) a. He is not **tall**. $\overset{\heartsuit+\spadesuit+\diamondsuit}{\Longrightarrow}$ He is **short**.

b. He is not **short**. $\overset{\heartsuit+\clubsuit+\diamondsuit}{\not\Rightarrow}$ He is **tall**.

Previous experimental investigation of the ITA

- Our experiments on LLMs **hugely rely on a previous study conducted on human subjects by Ruytenbeek et al., 2017**.
- In that study, the inference pattern presented in (1) was assessed in English and French using pairs like (2), whereby *too*, a presupposition trigger, forces synonymy between *not* $A^{\pm}$ in the first sentence and $A^{\mp}$ in the second sentence.

(2) a. John is not **tall**. Paul is **short** too.

b. # John is not **short**. Paul is **tall** too.

Refinement of Krifka's predictions: Ruytenbeek et al., 2017

- The experiment tested **morphologically opaque pairs** (like *tall*/*short*) and **morphologically transparent ones** (like *lucky*/*un-lucky*), in order to investigate a refinement of the previous theory, based on the following reasoning:
  - The decomposition $\textbf{A}^- = \textsc{Not}\text{-}\textbf{A}^+$ is made **particularly salient when the adjective is transparent**.
  - This means that (♠) and (♣) hold even "more unambiguously" for morphologically transparent pairs, which leads to a stronger interaction between the ITA and adjective polarity.
- In other words, **the contrast in (2) is expected to be stronger for transparent (T) antonyms (cf. (2′)) as opposed to opaque (O) ones**.

(2) a. John is not **tall**. Paul is **short** too.

b. # John is not **short**. Paul is **tall** too.

(2′) a. John is not **lucky**. Paul is **un-lucky** too.

b. ## John is not **unlucky**. Paul is **lucky** too.

Summary of the two semantic predictions at stake

- **H1** (Krifka, 2007, a.o.): it's easier to infer **A**$^{-}$ from *not* **A**$^{+}$ than vice-versa (**ITA strength** $\propto$ **adjective polarity**).
- **H2** (Ruytenbeek et al., 2017): the contrast in ITA strength is magnified with transparent adjectives (**ITA strength** $\propto$ **adjective polarity** $\times$ **morphological transparency**).

Technical and methodological background

Background: the Transformer architecture

- Recent LLMs are based on the **Transformer architecture** (Vaswani et al., 2017).
- A Transformer is a neural network whose basic building blocks involve **attentions mechanisms**. Attention mechanisms map the representation of a given token to a mixture of the representations of surrounding tokens, according to how "relevant" those tokens are to the current one.
- In practice, this allows LLMs to **model a diverse range of long-distance dependencies** within a sentence, and to assign each word (or rather, token), a **dynamic vector representation** which depends on its context.
- Transformers are primarily generative, which means that they will predict tokens given a context, by assigning them probabilities.
  - **Left-to-right models** (e.g. GPT family) compute token representations and probabilities in a left-to-right fashion;
  - **Bidirectional models** (e.g. BERT family) compute token representations and probabilities using both left and right contexts.

Evaluating LLMs on sentence acceptability judgment tasks

- 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 proxy for syntactic/semantic acceptability** when investigating the "linguistic" behavior of statistical models of language (E. Wilcox et al., 2018; Futrell et al., 2019; E. G. Wilcox et al., 2023).

$$\begin{aligned}\textsc{Acceptability}(w_t) &\simeq -\textsc{Surprisal}(w_t) \\ &= \log P(w_t | w_1 \dots w_{t-1})^3 \\ \textsc{Acceptability}(w_1 \dots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)^1\end{aligned}$$

- We will use the same kind of methodology here, in Task 1. Sentence-level and word-level surprisals were computed using the Python `minicons` library (Misra, 2022).

[1] In the case of BERT-like bidirectional models, this formula is adapted to masked language modeling: the probability of a word is computed given its left *and* right context.

Evaluating LLMs on logical inferences

- It is also possible to evaluate certain LLMs **on logical inferences directly**. That's what we will do in Task 2.
- In that case, the models at stake do not generate tokens, but instead are **fine-tuned to perform Natural Language Inference (NLI)**.
- From a pair of sentences, these models are able to output the strength ($\sim$probability) of the entailment (or contradiction) between them.
- Side note: because NLI is basically a classification task, the same pair of sentences may score very high for both entailment *and* contradiction!!

Overview of the study

- 111 pairs of English antonyms (48 T, 63 O) were manually created. Some redundancy in the adjectives used across pairs, due to synonymy.
- 3 main Tasks:
  - **"Behavioral"**: evaluate surprisal at the sentence- and word-level for minimal pairs following different kinds of templates ("*too*" paradigm, "meta" paradigm) to assess differences in ITA strength.
  - **"Inferential"**: ask LLMs fined-tuned for NLI to directly assess ITA strength.
  - **"Internal"**: compare LLMs' internal vector representations of antonyms and their respective negations to assess if contrasts in ITA strength translate into inherent topological differences.

Models tested

- 4 models were tested: GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019).
  - GPT-2 is purely left-to-right;
  - XLNet has a left-to-right architecture but its objective function allowed it to incorporate some bidirectional dependencies during training.
  - BERT and RoBERTa are bidirectional.
- The differing architectures of the LLMs influence the way they process and "judge" sentences, as well as individual words.

Task 1: “Behavioral” assessment based on surprisal measures

Testing "paradigms"

- 3 kinds of minimal pairs were assessed in 3 different sub-experiment. All pairs of sentences were counterbalanced for gender and "filled" the 111 possible (**A**$^+$, **A**$^-$) antonymic pairs.

(2′) "Postposed *too*" (very close to the stimuli in Ruytenbeek et al., 2017)

a. He is not **A**$^+$, and she is **A**$^-$ too.

b. # He is not **A**$^-$, and she is **A**$^+$ too.

(2″) "Anteposed *too*" (does more justice to left-to-right LLMs)

a. He is not **A**$^+$. She too is **A**$^-$.

b. # He is not **A**$^-$. She too is **A**$^+$.

(3) "Meta"

a. He is not **A**$^+$ means that he is **A**$^-$.

b. # He is not **A**$^-$ means that he is **A**$^+$.

- We focus on paradigm (2″) here (but see Appendix for the two others).

Anteposed too paradigm: results for H1 (sentence-level)

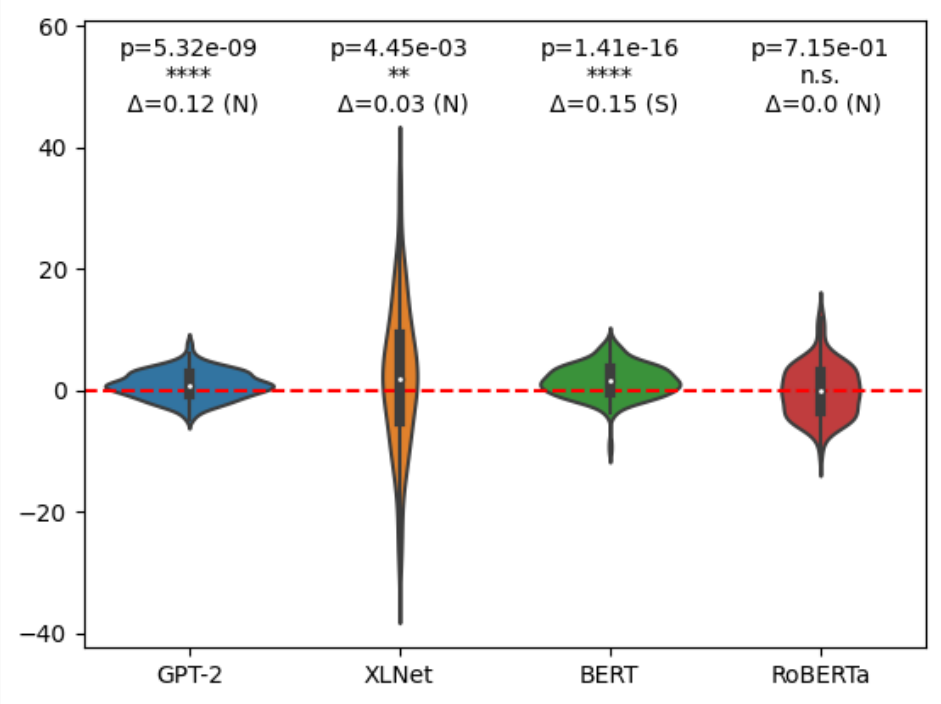


Figure 1: Paired differences in sentence surprisal between (2″b) and (2″a), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- All models but one (RoBERTa) exhibit a significant contrast in ITA strength as a function of adjective polarity, but the effect sizes are negligible (GPT-2/XLNet) or small (BERT).

Anteposed too paradigm: results for H2 (sentence-level)

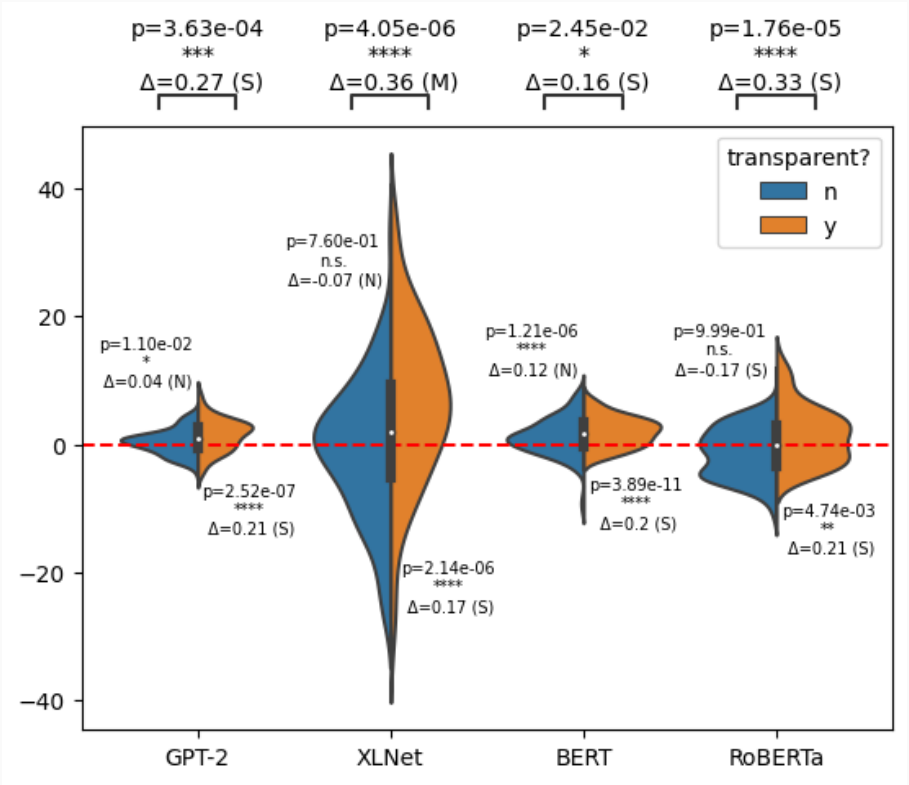


Figure 2: Paired differences in sentence surprisal between (2″b) and (2″a), group-by-group (T vs. O), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- GPT-2 and BERT are the two models for which H1 is individually verified by both the T- and O-group.
- Both models also verify H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group (small effect sizes).

Upshot of the sentence-level investigation

- **Weak evidence for the models (with the notable exception of XLNet) capturing the contrast in ITA strength regarding adjective polarity (H1) and the interaction with morphological transparency (H2).**
- No evidence for opposite contrasts, at least!!
- **But what do the best performing models do at the word-level?**
- From a language processing standpoint, we expect the positive contrasts in surprisal witnessed in the sentence-level assessments to be **driven by the occurrence of the second adjective**:
  - given what precedes it, this adjective is expected to be ok (i.e. not surprising) when **negative**;
  - and less ok (i.e. quite surprising) when **positive**.

(2″) a. He is not **A**$^{+}$. She too is **A**$^{-}_{\odot}$.

b. # He is not **A**$^{-}$. She too is **A**$^{+}_{\odot}$.

Word-level processing: GPT-2

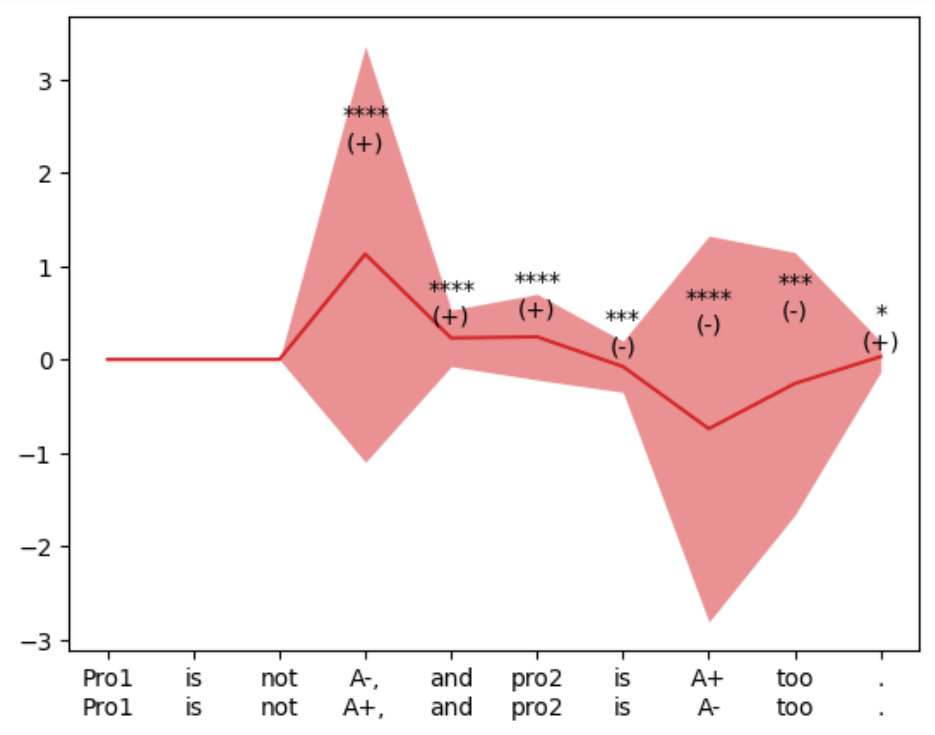


Figure 3: Paired word-by-word differences in surprisal between (2″b) and (2″a), *p*-values computed using Wilcoxon tests. Red line is the mean, red enveloppe is the standard deviation. Similar plots for the two other paradigms.

- $\mathbf{A}^-$ is significantly more surprising than $\mathbf{A}^+$ after negation (position 4)...
- but also in position 8 (second occurrence), against the expectations...
- **The effect witnessed at the sentence-level was driven by the wrong element of the sentence!!!**
- BERT and RoBERTa did better but evaluating bidirectional models at the word-level is also trickier (see Appendix for results).

Upshot of the behavioral Task

- Some LLMs seem to "behave" like human subjects at the sentence-level, although effect sizes are small.
- XLNet, which was supposed to combine the best aspects of left-to-right and bidirectional models (and beat GPT-2/BERT at "standard" NLP benchmarks), performed surprisingly poorly.
- Moreover, what the models do at the word-level does not always seem sensible.
- Before digging even further into the LLMs' representation of antonymic adjectives (Task 3), let's take a detour and try a more direct method of assessing ITA strength for stimuli sentences.

Task 2: “Inferential” assessment based on entailment probabilities

An even more direct way to assess ITA strength

- Recall the "meta" paradigm used in Task 1:

(3) a. He is not **A**$^+$ means that he is **A**$^-$.

b. # He is not **A**$^-$ means that he is **A**$^+$.

- We can test this contrast **even more directly, without appealing to surprisal measures**, by just asking a LLM fine-tuned to perform Natural Language Inference to output the probability of the following entailments (4) and contradictions (5):

(4) a. He is not **A**$^+$ $\Rightarrow$ He is **A**$^-$.

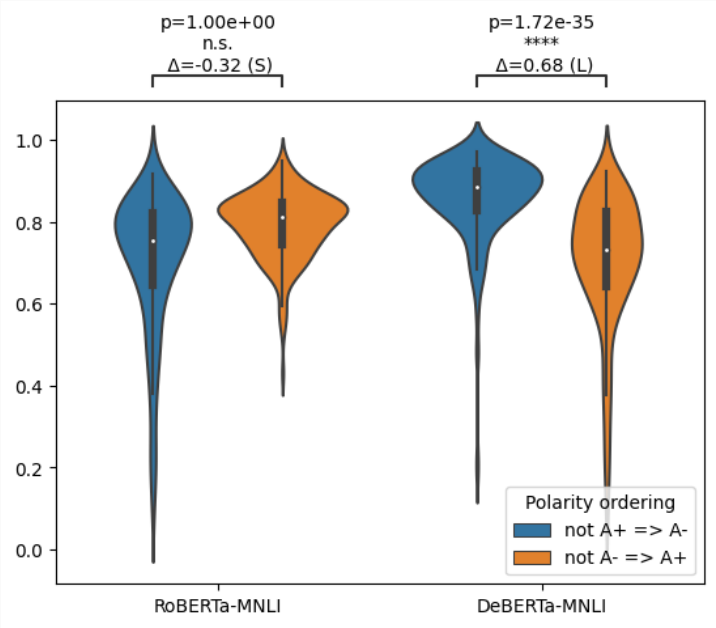
b. He is not **A**$^-$ $\Rightarrow$ He is **A**$^+$.

(5) a. He is not **A**$^+$ $\wedge$ He is **A**$^-$ $\Rightarrow \bot$.

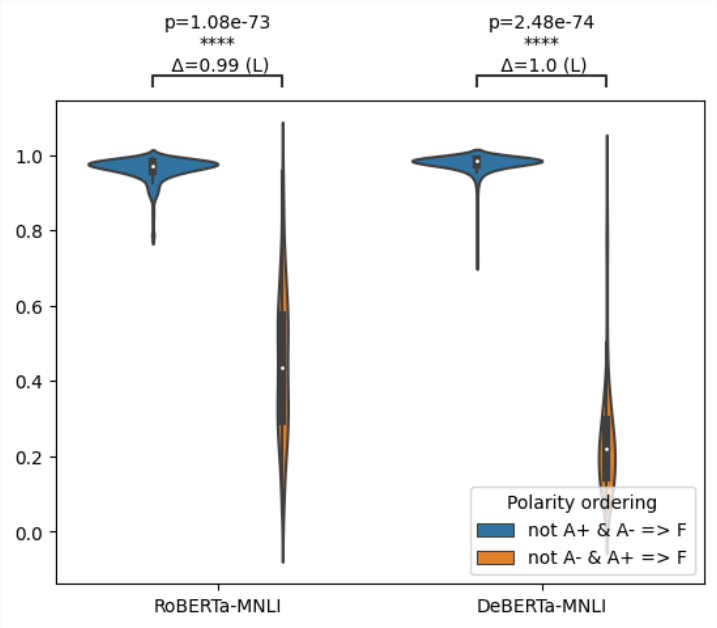
b. He is not **A**$^-$ $\wedge$ He is **A**$^+$ $\Rightarrow \bot$.

- **Based on H1, we predict the entailment in (4a) to have a higher probability than the one in (4b) and the contradiction in (5a) to have a lower probability than the one in (5b).**

Results (H1 only)



(a) Probabilities of entailment for (4a) and (4b), and two LLMs fined-tuned for NLI.



(b) Probabilities of contradiction for (5a) and (5b), and two LLMs fined-tuned for NLI.

- Regarding **entailment predictions**, only one model (DeBERTa) correctly predicts the inference (4a) to be stronger than (4b)...
- Regarding **contradiction predictions**, *both* models wrongly predict (5a) to be stronger than (5b), with very high confidence...

Task 3: “Internal” assessment based on vector similarities

Core idea

- In this task, we abandon stimuli sentences to focus on the **internal (vector) representations assigned by the original standard LLMs to $\mathbf{A}^+$, $\mathbf{A}^-$, and their respective negations**: $\overrightarrow{\mathbf{A}^+}$, $\overrightarrow{\mathbf{A}^-}$, $\overrightarrow{\text{not } \mathbf{A}^+}$, $\overrightarrow{\text{not } \mathbf{A}^-}$.[2]
- A common measure of semantic proximity in such vector spaces is cosine similarity:

$$CosSim(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 . \vec{v}_2}{||\vec{v}_1|| \times ||\vec{v}_2||} \in [-1; 1]$$

- If H1 translates into the LLMs' vector space, we then expect $\overrightarrow{\text{not } \mathbf{A}^+}$ to be closer to $\overrightarrow{\mathbf{A}^-}$ than $\overrightarrow{\text{not } \mathbf{A}^-}$ is close to $\overrightarrow{\mathbf{A}^+}$, i.e.:

$$CosSim(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-}) - CosSim(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+}) > 0$$

- Moreover, H2 predicts that this difference should be bigger for T-antonyms as opposed to O-antonyms.

[2] In practice, we included the copula *is* as a left context to get those representations.

Results for H1, both groups

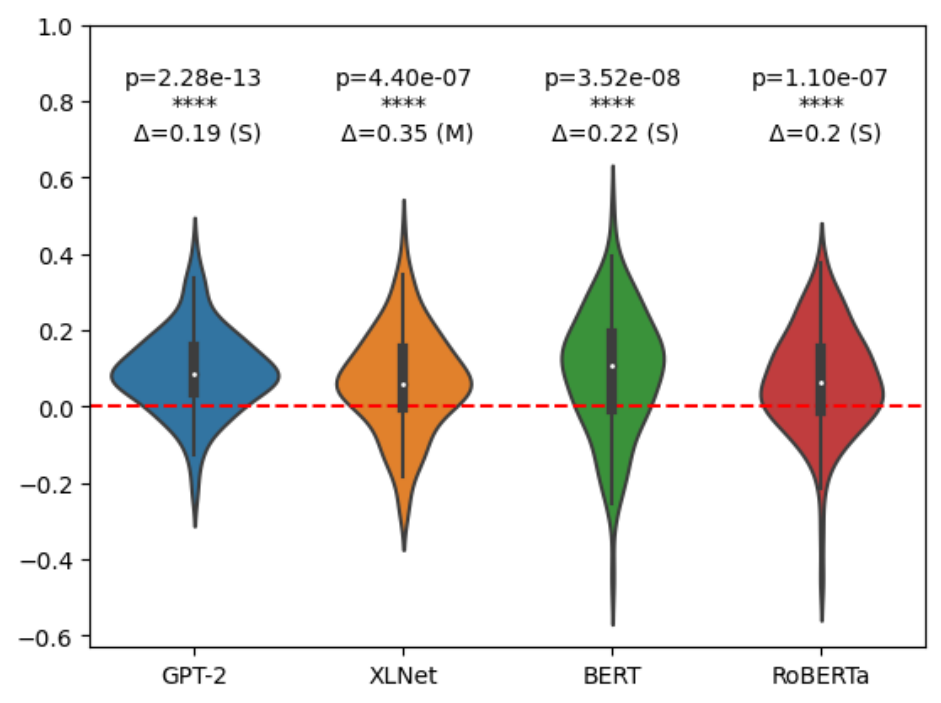


Figure 5: Paired differences in cosine similarities between $(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-})$ and $(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+})$, p-value computed using a Wilcoxon test, effect sizes using Cliff's Δ.

- All models exhibit a **significant contrast in cosine similarities (and by proxy ITA strength) as a function of adjective polarity**, with small-to-medium effect sizes.
- This suggests that H1 translates into a topological inequality within the LLMs' vector spaces!

Results for H1, group-by-group, and H2

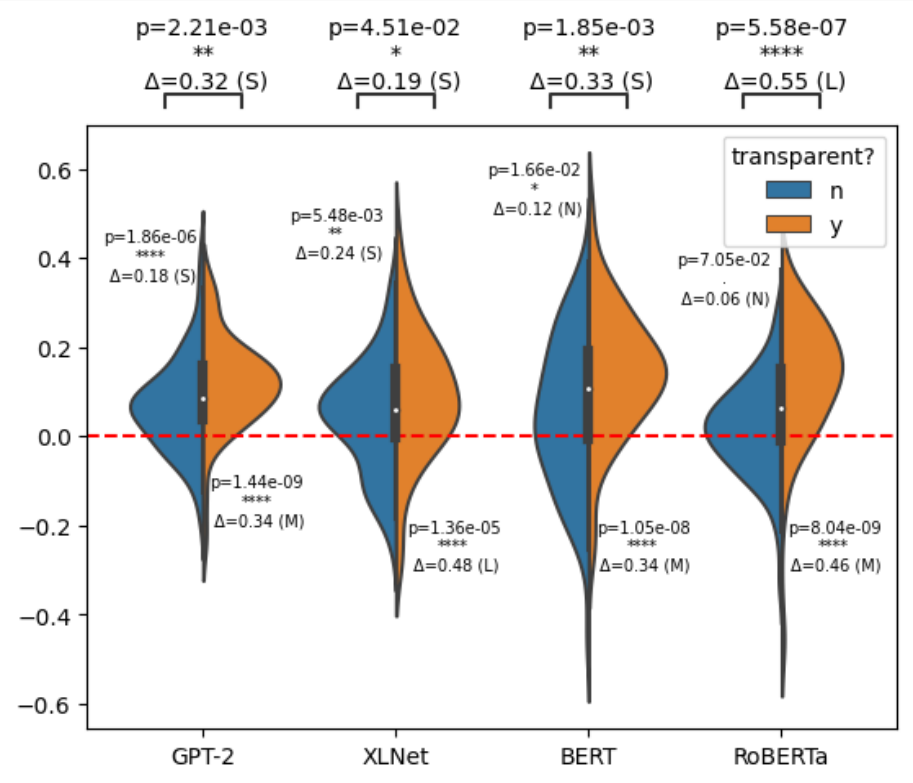
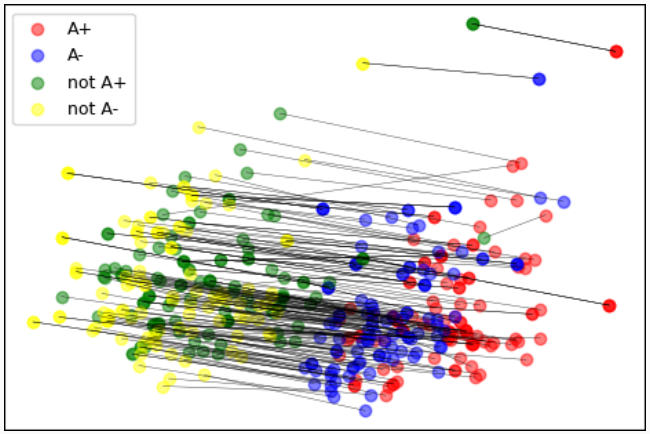


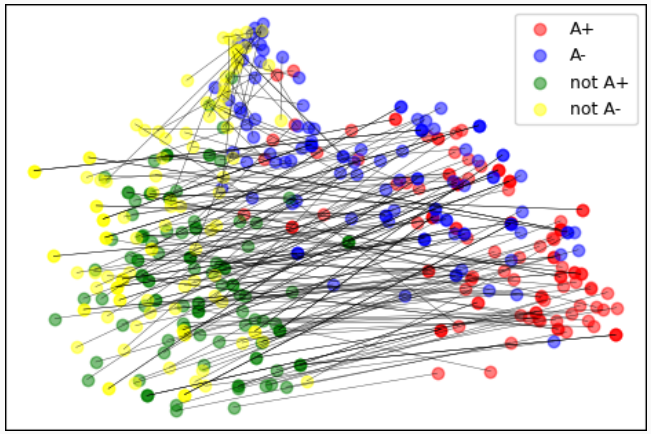
Figure 6: Paired differences in cosine similarities between $(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-})$ and $(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+})$, group-by-group p-values computed using a Wilcoxon test, and between-group p-values using a Mann-Whitney U-test. Effect sizes are Cliff's Δ.

- GPT-2 and XLNet are the two models for which H1 is individually verified by both the T- and O-group.
- Both models also verify H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group (small effect sizes).
- **Quite encouraging results overall but...**

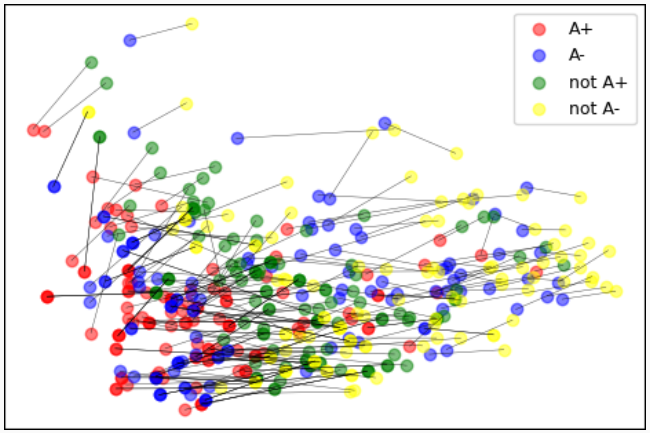
The big picture: fairly worrying



(a) GPT-2



(b) XLNet



(c) BERT

- The bare antonyms (blue and red dots) on the one hand, and their negations (yellow and green dots) on the other hand, **cluster together** in a reduced 2D space!!!
- This effect is evidently way bigger than the one measured previously, and **replicates the negative result of Aina et al., 2019** for earlier models.

Conclusion

Upshot of the 3 tasks

- Although the target predictions regarding the ITA (H1, H2) were more or less verified across tasks, as soon as we dig a little bit deeper, reasonable expectations about the models's behavior are not met:
  - In Task 1, LLMs manage to give human-like "judgments" on minimal pairs involving (negated) antonyms, but **do not seem to focus on the right individual words** to produce them.
  - In Task 2, one LLM performing Natural Language Inference seem to capture the expected interaction between adjective polarity and inference strength, but despite this **it does not differentiate at all between inference types (entailment vs. contradiction)**.
  - In Task 3, LLMs manage to translate some contrast in semantic similarity between negated adjectives and their antonym in their internal vector space, but **those spaces are characterized by stronger, very much unexpected topological regularities**.

Outlook

- **Methodological**: it is not enough to only test the critical predictions of a given theory on a given LLM. One should always keep in mind the big picture, and assume the model does *not* share any common sense assumptions with us.
- **Theoretical**: antonyms and their negation are acquired relatively easily (although it is true that the ITA may be trickier...). But the fact that LLMs did not manage to fully imitate human capacity in that domain despite being exposed to many instances of the relevant adjectives in various contexts, may suggest two things:
  - Either the acquisition of antonyms and of the ITA requires more grounding (anchoring to the actual world) than what the LLMs under study had access to.
  - Or, human cognition (as opposed to purely statistical learning) comes with a few useful biases (e.g. the fact that complexity/markedness influences pragmatic inferences) making the learning of antonyms easier.

Thank you !

Selected references i

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Charles, W. G., & Miller, G. A. (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics*, *10*(3), 357–375. https://doi.org/10.1017/S0142716400008675

Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press.

Justeson, J. S., & Katz, S. M. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, *17*(1), 1–20. https://aclanthology.org/J91-1001

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. https://doi.org/10.3115/1073336.1073357

Büring, D. (2007a). Cross-polar nomalies. *Semantics and Linguistic Theory*, *17*, 37. https://doi.org/10.3765/salt.v17i0.2957

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

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 163–177). Palgrave Macmillan UK. https://doi.org/10.1057/9780230210752_6

Selected references ii

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: a journal of general linguistics*, *2*(1). https://doi.org/10.5334/gjgl.151

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

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

Gotzner, N., Solt, S., & Benz, A. (2018). Adjectival scales and three types of implicature. *Semantics and Linguistic Theory*, *28*, 409. https://doi.org/10.3765/salt.v28i0.4445

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. https://doi.org/10.18653/v1/W18-5423

Aina, L., Bernardi, R., & Fernández, R. (2019). Negated adjectives and antonyms in distributional semantics: Not similar? *Italian Journal of Computational Linguistics*, *5*(1), 57–71. https://doi.org/10.4000/ijcol.457

Selected references iii

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *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)*, 32–42. https://doi.org/10.18653/v1/N19-1004

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

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

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & 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). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

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

Selected references iv

Wilcox, E. G., Futrell, R., & Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44. https://doi.org/10.1162/ling_a_00491

Appendices

Some extra background on positive and negative adjectives

- It has been observed that intuitively positive vs. negative adjectives pattern differently in several respects...
  - Positive (rather than negative) adjectives are used to **ask unbiased degree-related questions**.
  - Positive (rather than negative) adjectives are used to **form unbiased comparatives/equatives**.
  - Negative (rather than positive) adjectives may **feature overt negative morphology**.

(6) a. How tall is John? $\rightsquigarrow$ John may be tall or short.

b. How short is John? $\rightsquigarrow$ John is short.

(7) a. John is as tall as Paul. $\rightsquigarrow$ Both may be tall or short.

b. John is as short as Paul. $\rightsquigarrow$ Both are short.

(8) a. in-competent; im-modest; un-lucky; dis-honest ...

b. *un-small; *im-messy; *un-poor; *dis-arrogant ...

Postposed too paradigm at the sentence-level: both groups

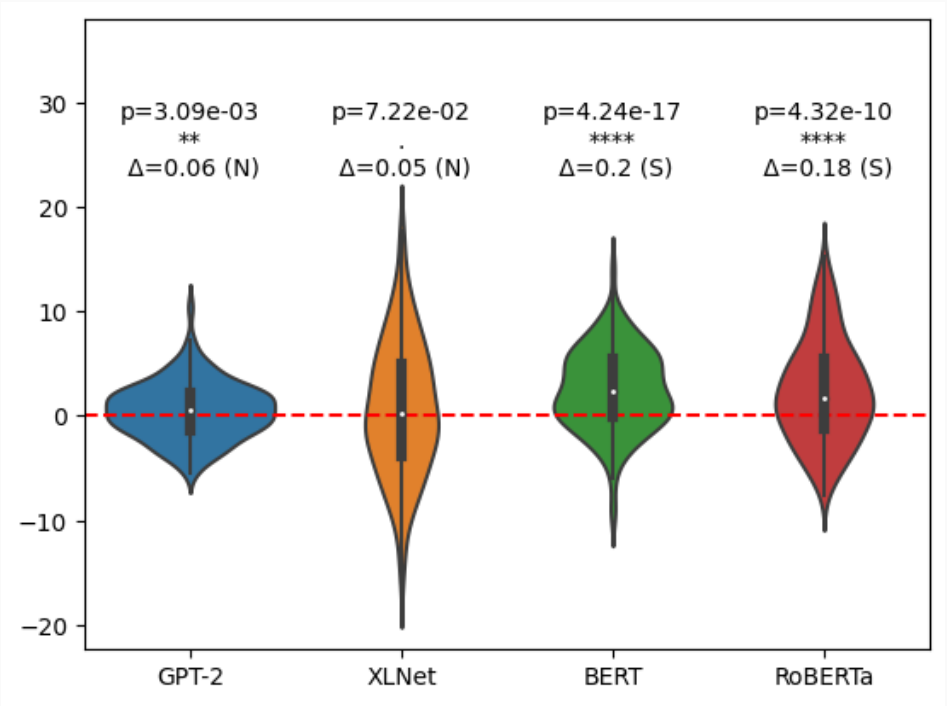


Figure 8: Paired differences in sentence surprisal between (5′b) and (5′a), p-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- All models but one (XLNet) exhibit a significant contrast in ITA strength, but the effect sizes are negligible (GPT-2) or small (BERT/RoBERTa).
- Because *too* appears after the critical adjectives, this paradigm expectedly favors bidirectional models.

Postposed too paradigm at the sentence-level: group-by-group

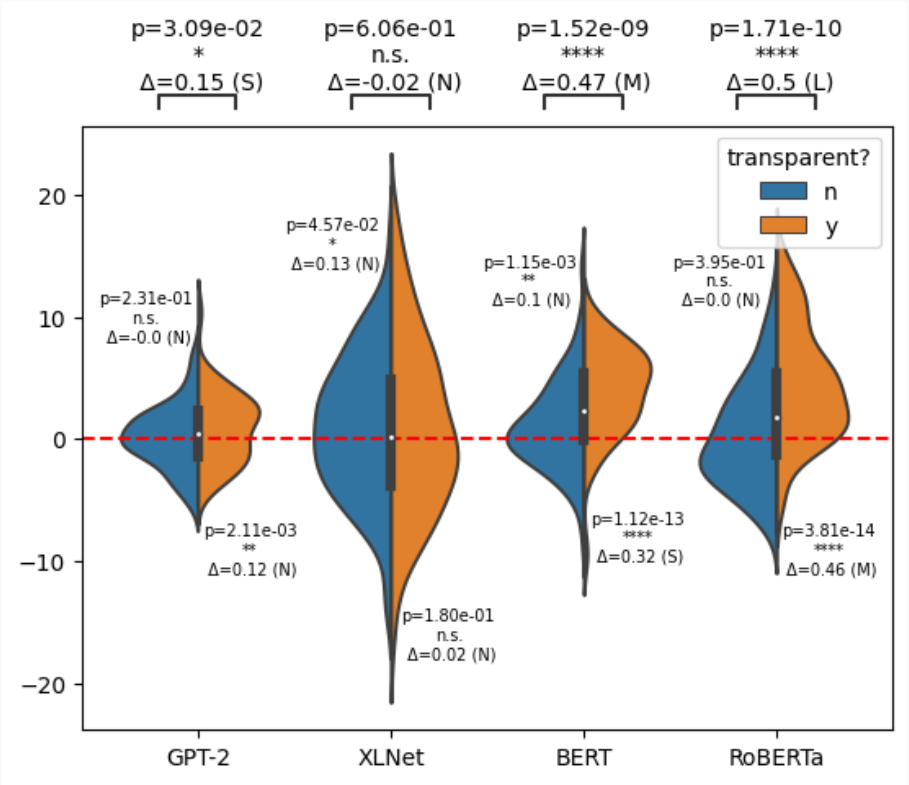


Figure 9: Paired differences in sentence surprisal between (5′b) and (5′a), group-by-group (T vs. O), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- BERT is the only model for which H1 is individually verified by both the T- and O-group.
- BERT also verifies H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group (medium effect size).

"Meta" paradigm at the sentence-level:both groups

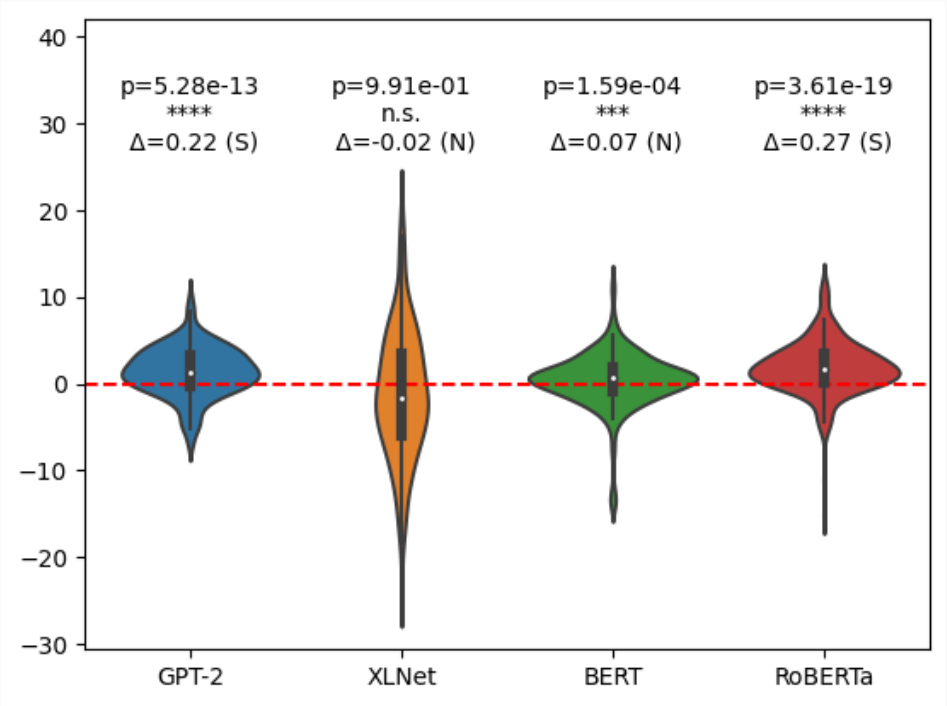


Figure 10: Paired differences in sentence surprisal between (3b) and (3a), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- All models but one (XLNet) exhibit a significant contrast in ITA strength, but the effect sizes are negligible (BERT) or small (GPT-2/RoBERTa).

"Meta" paradigm at the sentence-level:group-by-group

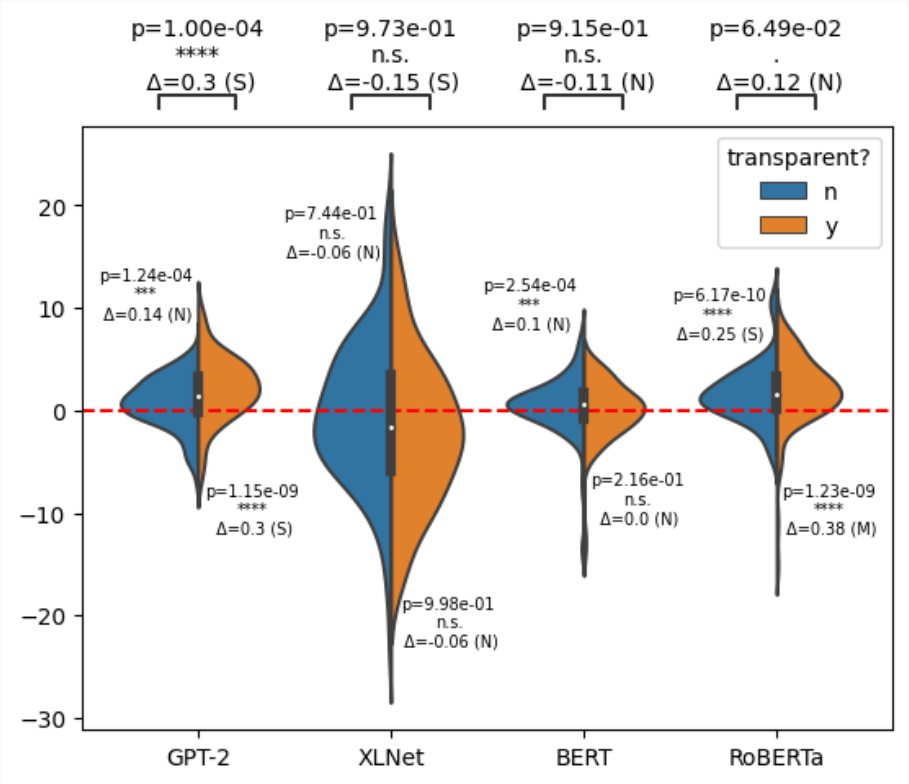


Figure 11: Paired differences in sentence surprisal between (3b) and (3a), group-by-group (T vs. O), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- GPT-2 and RoBERTa are the two models for which H1 is individually verified by both the T- and O-group.
- But only GPT-2 clearly verifies H2 (RoBERTa is characterized by a negligible effect size...).

Word-level processing: BERT

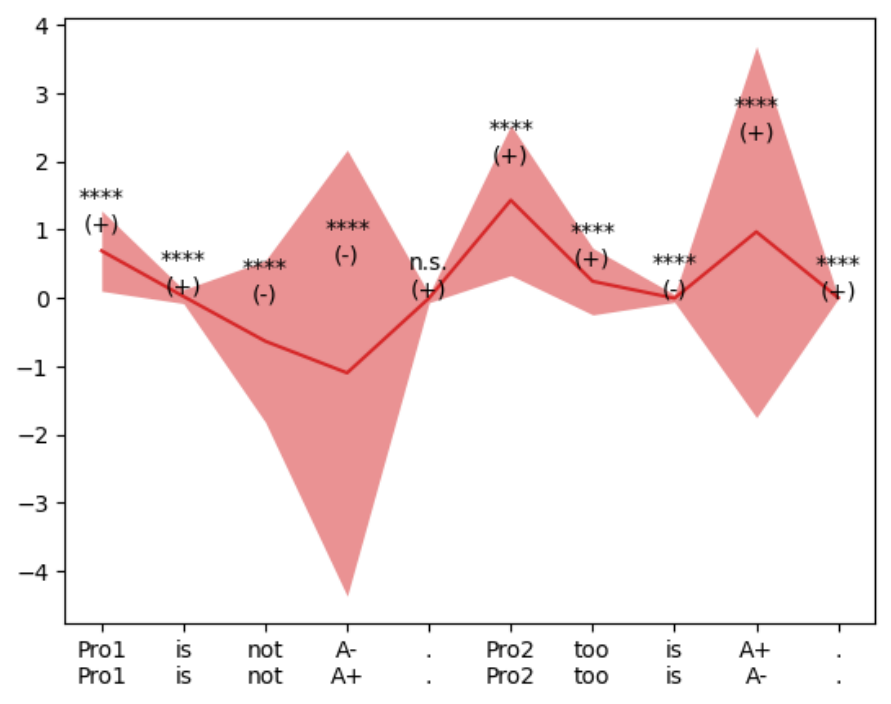


Figure 12: Paired word-by-word differences in surprisal between (3b) and (3a), *p*-values computed using Wilcoxon tests. Red line is the mean, red enveloppe is the standard deviation. Similar plots for the two other paradigms.

- $\mathbf{A}^-$ is significantly less surprising than $\mathbf{A}^+$ after negation (position 4)...
- and also significantly less surprising than $\mathbf{A}^+$ in position 9.
- The effect witnessed at the sentence-level makes sense at the word-level.
- But some amount of negative surprisal may have "transferred" from position 9 to position 4, due to the model's bidirectionality.