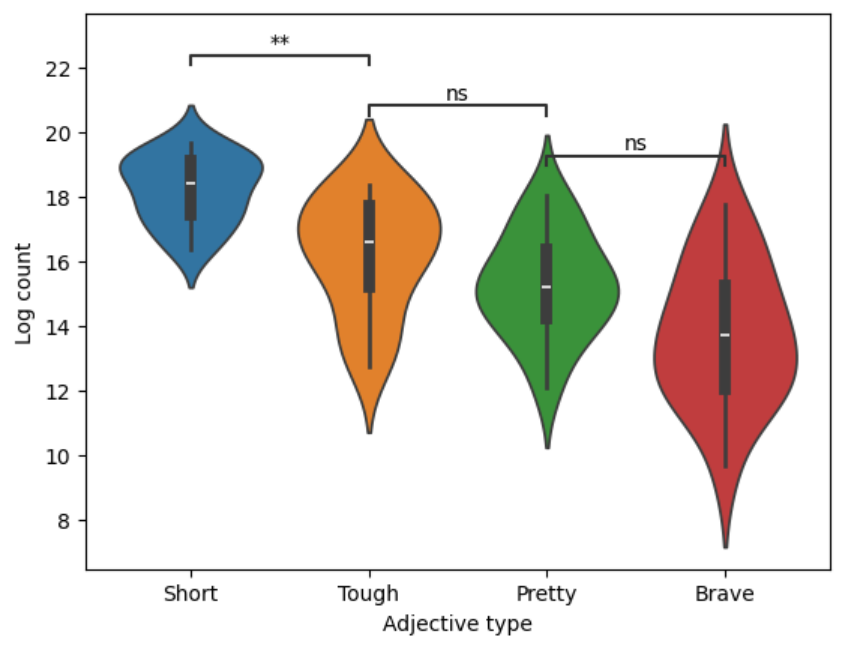
Syntax and grounding in adjective learning
Supplementary material for Study 1

Adèle Hénot-Mortier (MIT)

February 27, 2025

Queen Mary University of London

Adjective frequencies



Distributions of log counts, Mann-Whitney U tests, BY-corrected.

- Counts extracted from a Kaggle dataset based on the Google Web Trillion Word Corpus.
- **Short**-like adjectives from our dataset are significantly more common than the others.
- No difference between the other groups (small sample sizes, $n = 16...$), though a trend is visible.

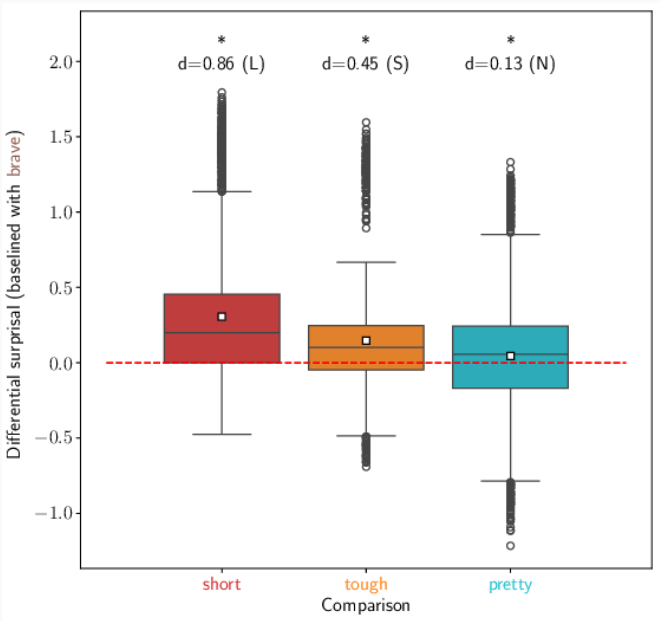
- Even if **short**-adjectives may be expected to be **less "surprising"** for LLMs, we will see they turn out surprising if put in the **wrong** syntactic environment.
- Even if **brave**-adjectives may be expected to be **more "surprising"**, they turn out unsurprising if put in the **right** syntactic environment.

GPT-2 Surprisal Assessment

GPT-2 + "of to" template (supporting brave)

(1) It's **A** of **X** to **V** this **Y**. ✓B; ✗S; ✗T; ✗P

Example: It's **brave** of **you** to **jister** this **kress**.



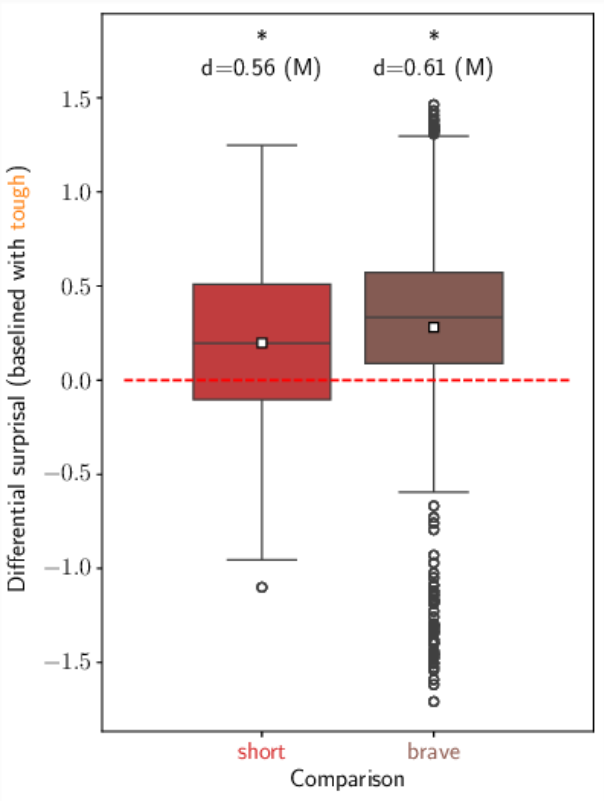
GPT-2 Large.
p-values are BY-corrected. * means $p < .05$. Effect sizes are Cohen's *d*.
N=Negligible, S=Small, L=Large.

- Predictions:
  - $\mathscr{S}$((1)+**short**) - $\mathscr{S}$((1)+**brave**) > 0
  - $\mathscr{S}$((1)+**tough**) - $\mathscr{S}$((1)+**brave**) > 0
  - $\mathscr{S}$((1)+**pretty**) - $\mathscr{S}$((1)+**brave**) > 0
- **Brave-adjectives are close to being the least surprising**: less surprising than **short** and **tough**, and on a par with **pretty**-like predicates.

GPT-2 + "bare to" template (supporting tough and pretty)

(2) This **Y** is **A** to **V**. ✓**T**; ✓**P**; ??**B**; ✗**S**

Example: This **quirm** is **tough** to **scarpe**.



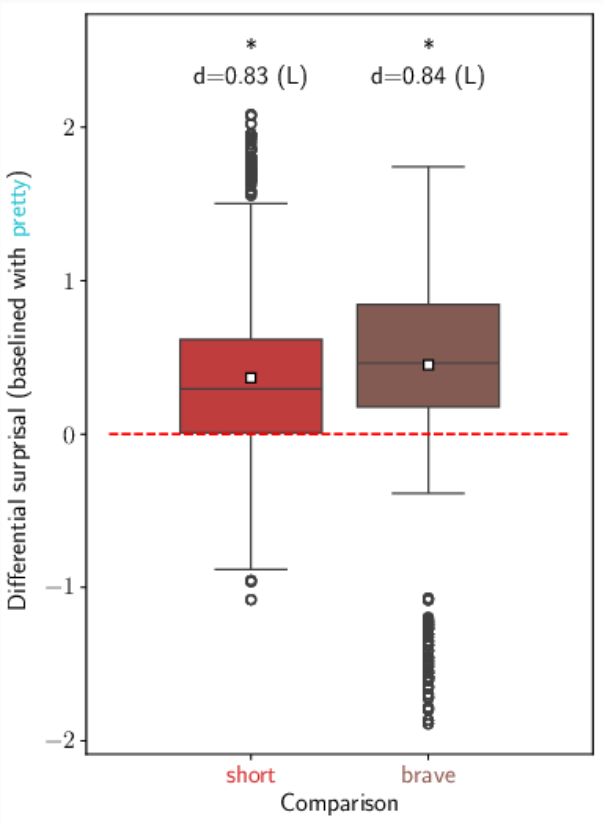
GPT-2 Large.
White squares are means. p-values are BY-corrected.
* means $p < .05$. Effect sizes are Cohen's d.
M=Medium.

- Predictions for **tough** vs. **brave** and **short**:
  - $\mathscr{S}((1)+\textbf{short})$ - $\mathscr{S}((1)+\textbf{tough}) > 0$
  - $\mathscr{S}((1)+\textbf{brave})$ - $\mathscr{S}((1)+\textbf{tough}) > 0$
- **tough**-adjectives are less surprising than **short** and **brave**-adjectives, in line with intuitions, though the **brave**-plot being more positive than the **short**-plot is a bit surprising, cf. caveat slide.

GPT-2 + "bare to" template (supporting tough and pretty)

(2) This **Y** is **A** to **V**. ✓P; ✓T; ??B; ✗S

Example: This **quirm** is pretty to **scarpe**.



GPT-2 Large.
White squares are means. p-values are BY-corrected.
* means $p < .05$. Effect sizes are Cohen's d. L=Large.

- Predictions for pretty vs. brave and short:
  - $\mathscr{S}$((1)+short) - $\mathscr{S}$((1)+pretty) > 0
  - $\mathscr{S}$((1)+brave) - $\mathscr{S}$((1)+pretty) > 0
- pretty-adjectives are less surprising than short and brave-adjectives, in line with intuitions, though the brave-plot being more positive than the short-plot is a bit surprising, cf. caveat slide.

Caveat with the "bare to" template (2)

- (2) features an objectless nonce verb in the infinitival clause.

(2) This **Y** is **A** to **V**. ✓**T**; ✓**P**; ??**B**; ✗**S**

Example: This **quirm** is **tasty** to **scarpe**.

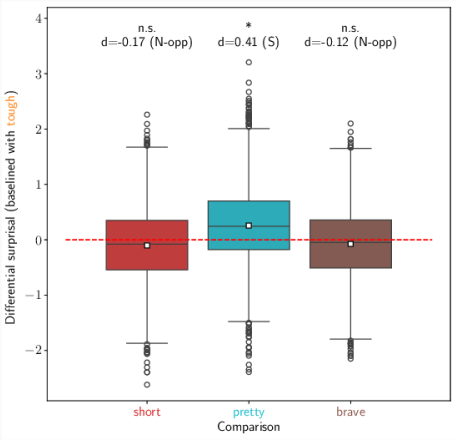
??This **quirm** is **rude** to **scarpe**.

- This verb could be understood as intransitive, or transitive.
  - If transitive, it needs a gap, and the whole construction supports matrix **tough**- and **pretty**-adjectives.
  - If intransitive, it does not need a gap, and the whole construction supports matrix **brave**-adjectives.
- Given that transitive verbs are more common than intransitive ones,[1] we expect (2) to favor **tough**- and **pretty**-adjectives, against **short**- and **brave**-adjectives.
- The previous slide shows this is the case, although **brave**-adjectives seem surprisingly *more dispreferred* than **short**-adjectives.

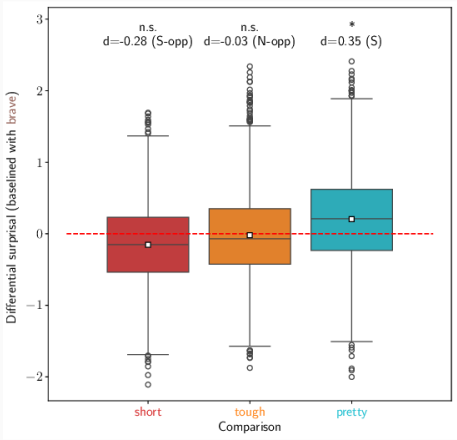
[1] Although this does not directly reflect frequencies, there are 9,778 "intransitive verb" entries on Wikipedia, and 21,273 "transitive verb" entries.

Surprisal assessment for other models

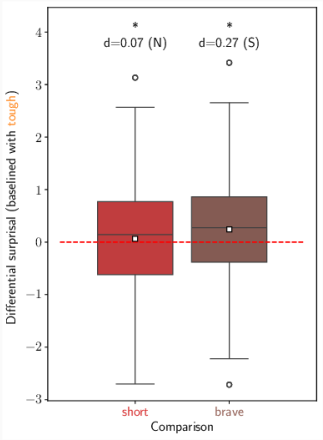
XLNet Large Surprisal assessement

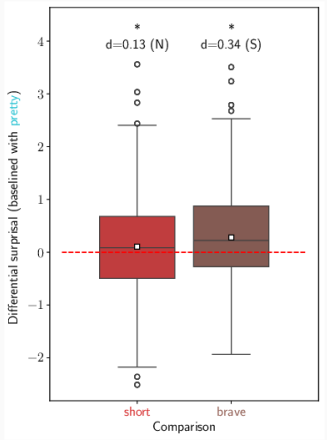


(a) "*for-to*"



(b) "*of-to*"



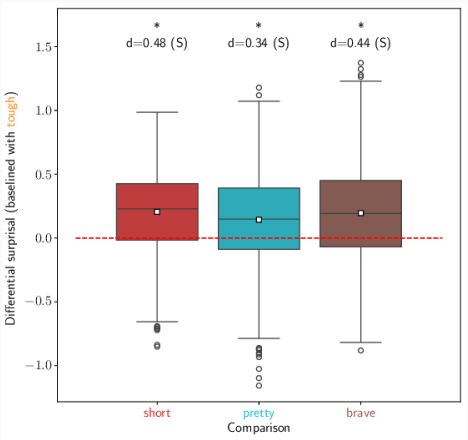


(c) "Bare *to*"

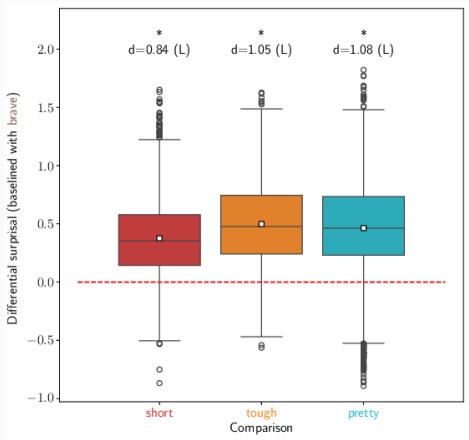
XLNet Large. Same tests and conventions as before. the -opp suffix on effect sizes mean the effect goes in the opposite direction.

- Poor results overall, except for predictions involving the **pretty**-class.
- Interesting, given that XLNet was shown to outperform GPT-2 on standard benchmarks!

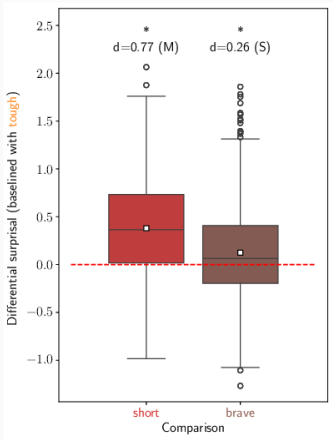
BERT Large Surprisal assessment

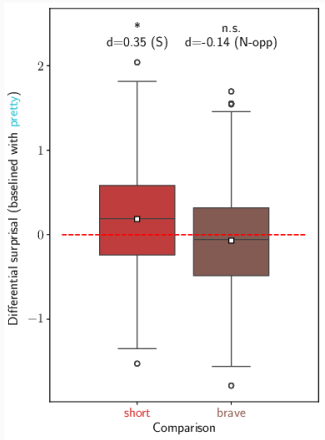


(a) "*for-to*"



(b) "*of-to*"



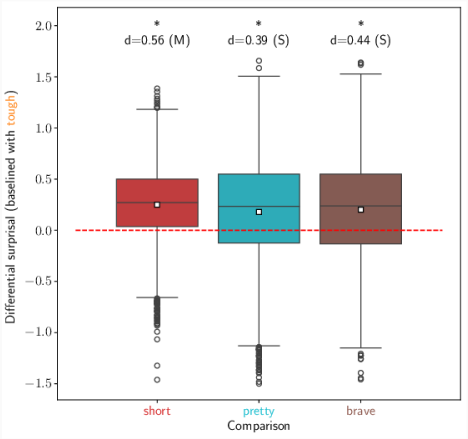


(c) "Bare *to*"

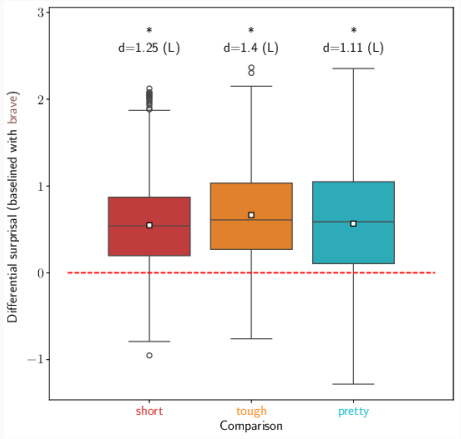
BERT Large. Same tests and conventions as before. the -opp suffix on effect sizes mean the effect goes in the opposite direction.

- Pretty good results overall, especially for the *of-to* template supporting **brave**-adjectives.

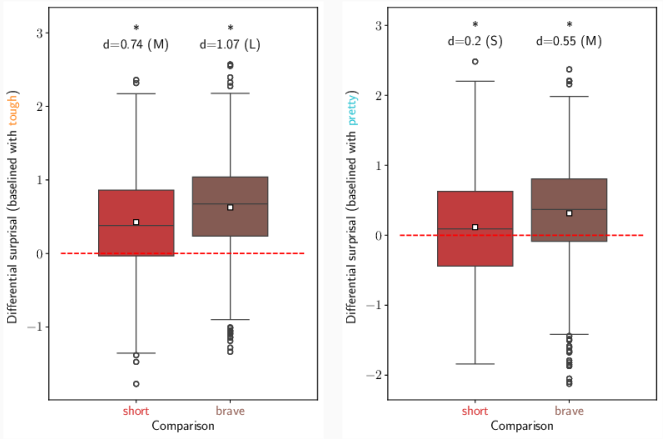
RoBERTa Large Surprisal assessment



(a) "*for-to*"



(b) "*of-to*"

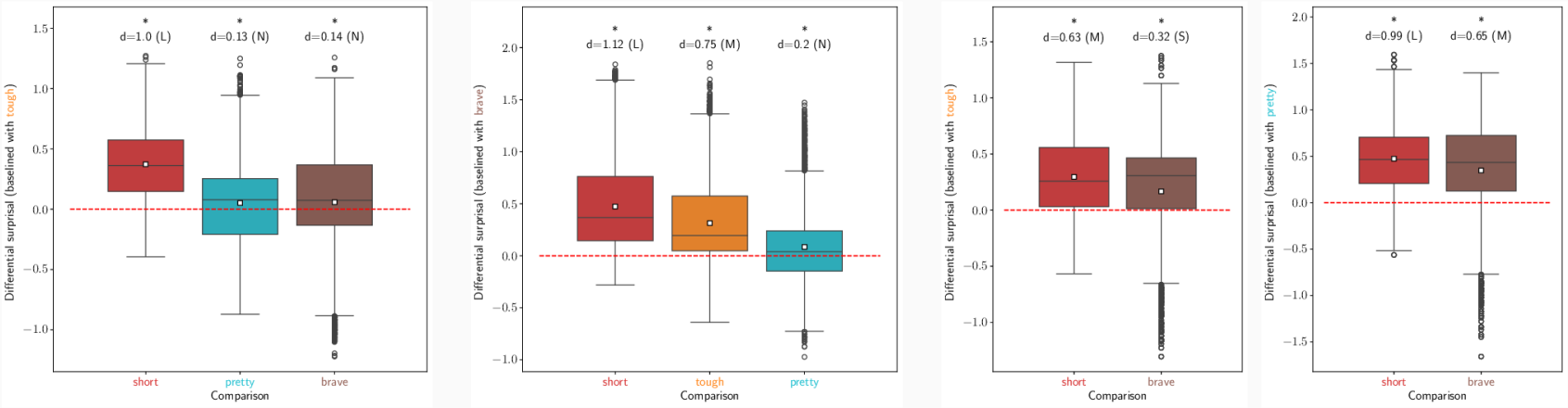


(c) "Bare *to*"

RoBERTa Large. Same tests and conventions as before. the -opp suffix on effect sizes mean the effect goes in the opposite direction.

- Very good results overall, especially for the *of-to* template supporting **brave**-adjectives.
- Some improvement w.r.t. RoBERTa's predecessor BERT!

Mistral 7B Surprisal assessment



(a) "*for-to*" **(b)** "*of-to*" **(c)** "Bare *to*"

Mistral 7B v0.1. Same tests and conventions as before. the -opp suffix on effect sizes mean the effect goes in the opposite direction.

- Mixed results. Quite inconclusive in the case of the *for-to* template supporting **tough**-like adjectives; better with the *of-to* and bare *to* templates.
- Interesting that the results appear less good than with GPT-2, BERT, or RoBERTa: Mistral 7B is a newer model, better on most if not all standard benchmarks.

Neural assessment

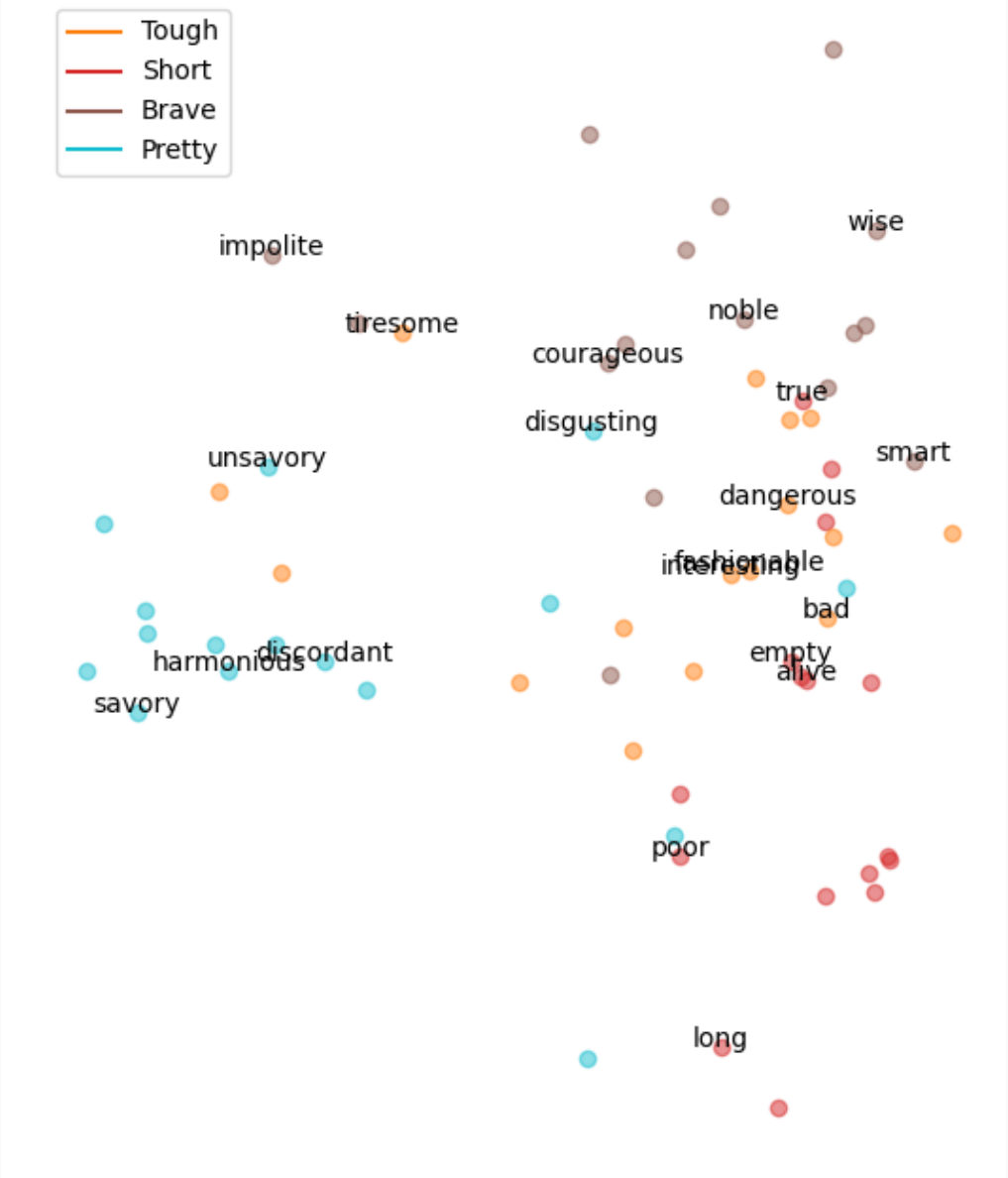
Embedding specifics

- We extracted contextualized vectors representations from the penultimate layer of all models (except Mistral).[2]
- The context provided for each adjective was: *It is* **A**.
- The penultimate layer is often argued to be contextualized enough, while remaining not too task-specific (i.e. it's not just aiming to predict a token).
- The higher dimension used to quantitatively assess clustering quality was obtained by performing a PCA set up to retain 90% of the explained variance.
- For GPT-2, this led to a space of dimension 39; for XLNet, 40; for BERT, 43, for RoBERTa, 42.

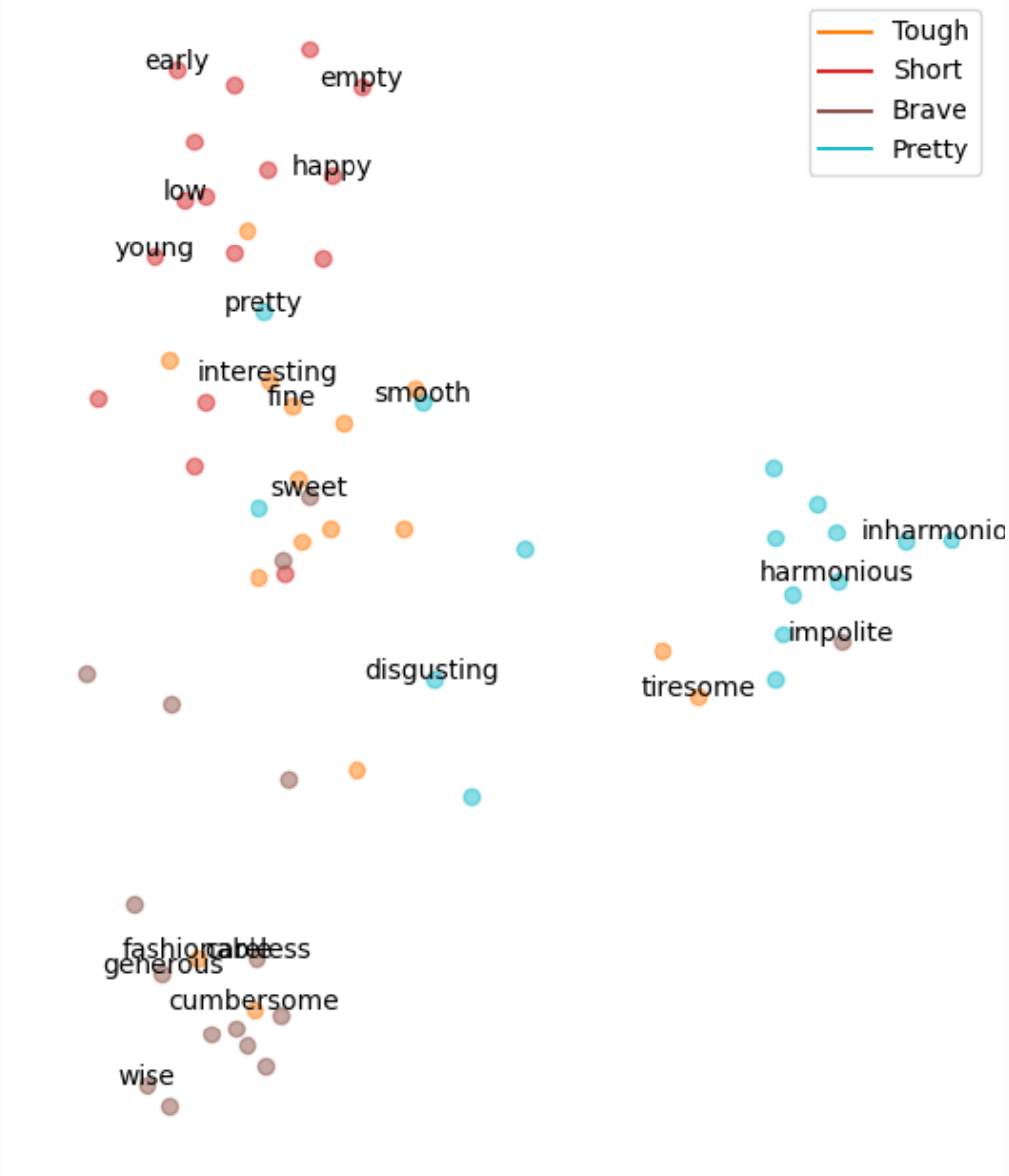
[2]We tried with the last layer as well; the clustering was of comparable or lesser quality in higher dimensions, and led to less clear 2D reductions.

XLNet Large vector space (2D)

BERT Large vector space (2D)



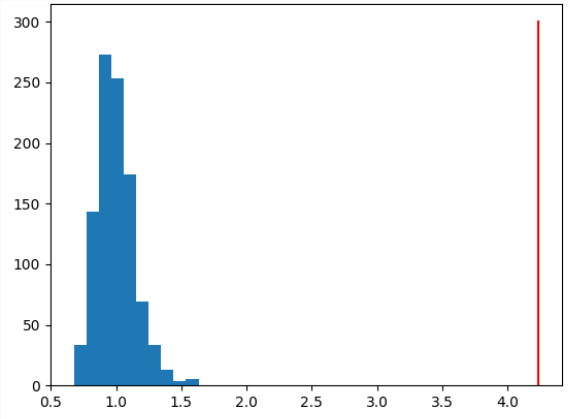
RoBERTa Large vector space (2D)



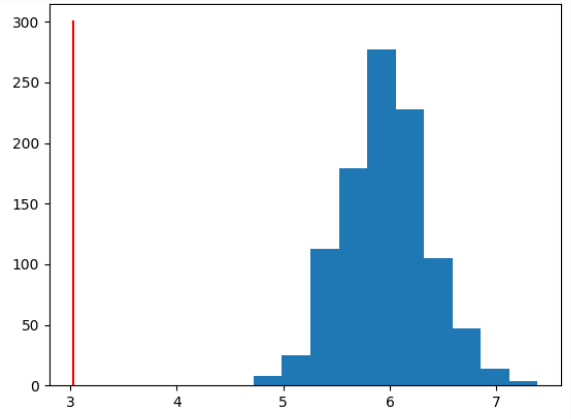
Assessing clustering quality in higher dimensions

- To quantitatively test if clustering was maintained in higher dimensions, I compared 4 empirical, "ground truth" clustering quality scores to the same scores obtained after randomly reassigning adjectives to vectors.
- Under the null hypothesis that models do not distinguish between adjective classes, randomly shuffling the vectors should not affect clustering quality – which should be pretty bad.
- The empirical scores ended up being clearly out-of-distribution, always in the good sense, for all models. In other words, LLMs assign adjectives to contextualized vector representations that encode their class (**tough**, **pretty**, **brave**, or **short**) – even if the clustering is not always readily visible in 2D.

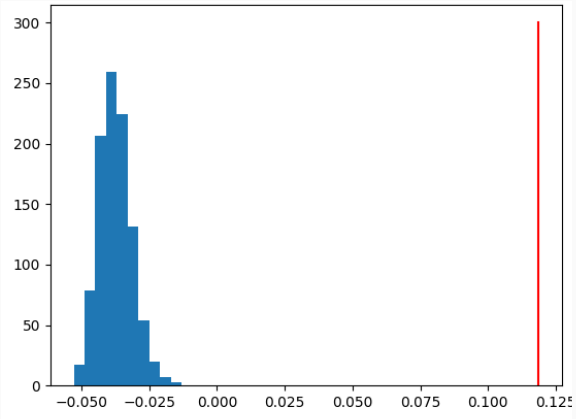
Clustering permutation plots: GPT-2 (d=39)



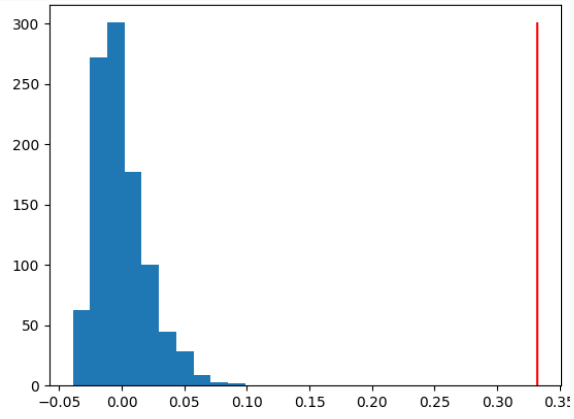
(a) Calinski-Harabasz (higher-better)



(b) Davies-Bouldin (lower-better)



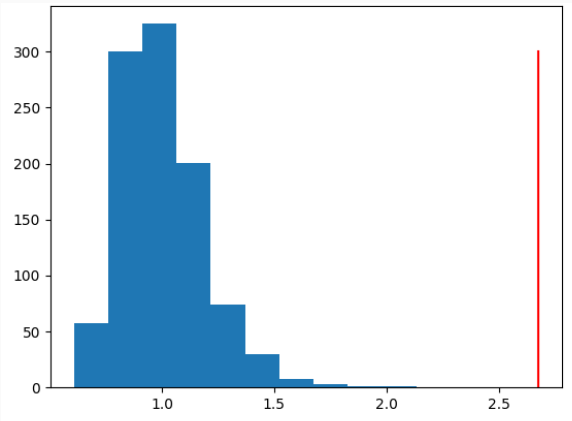
(c) Silhouette (higher-better)



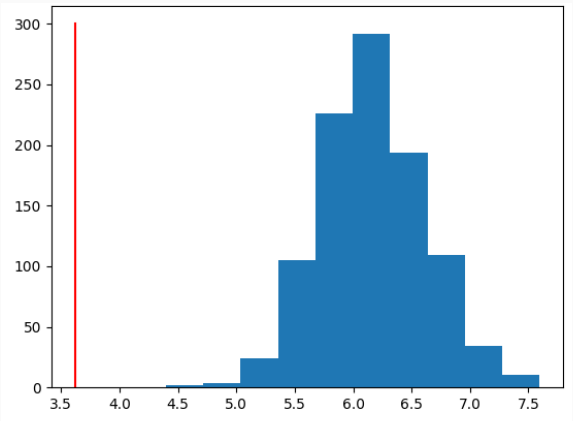
(d) Adjusted Rand (higher-better)

Distribution of quality scores for 1000 random labelings (blue histogram) vs. ground-truth scores (red line).

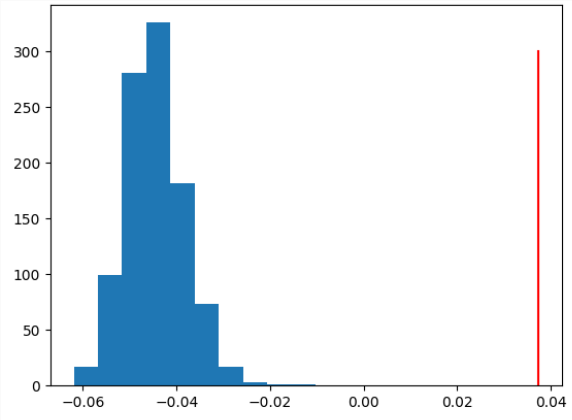
Clustering permutation plots: XLNet (d=40)



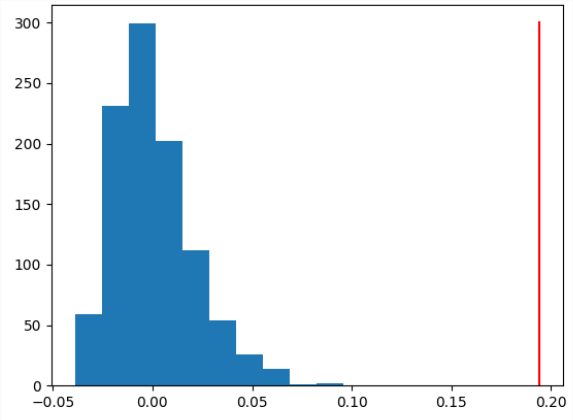
(a) Calinski-Harabasz (higher-better)



(b) Davies-Bouldin (lower-better)



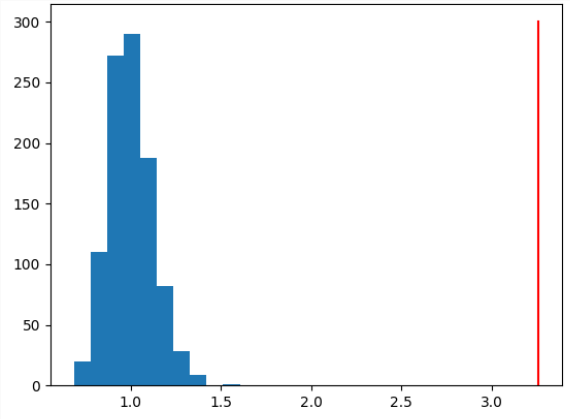
(c) Silhouette (higher-better)



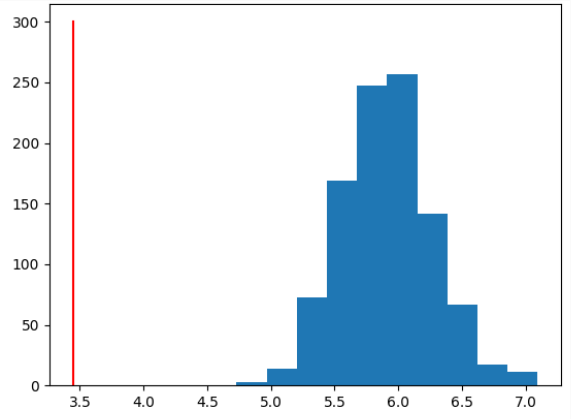
(d) Adjusted Rand (higher-better)

Distribution of quality scores for 1000 random labelings (blue histogram) vs. ground-truth scores (red line).

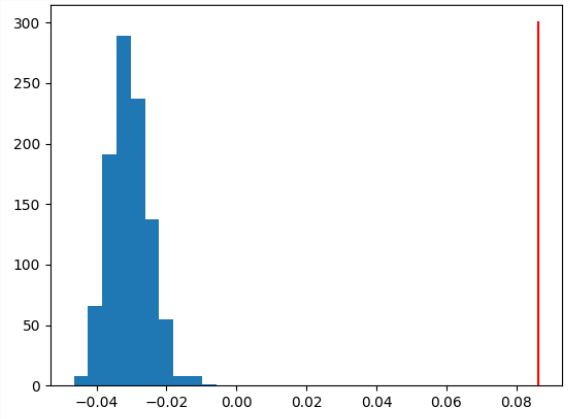
Clustering permutation plots: BERT (d=43)



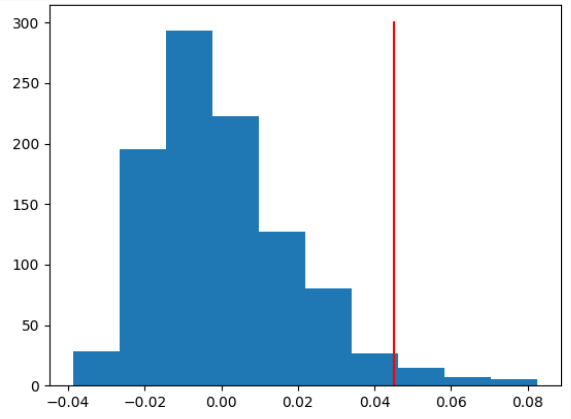
(a) Calinski-Harabasz (higher-better)



(b) Davies-Bouldin (lower-better)



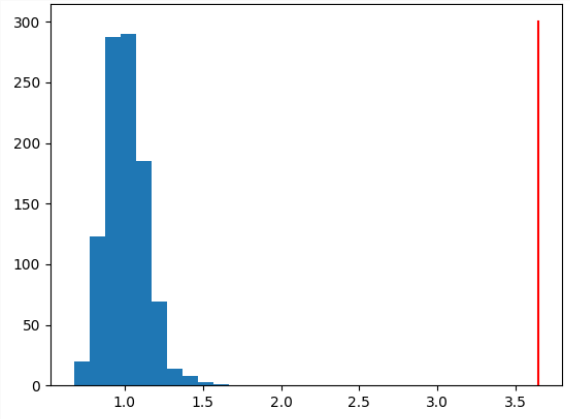
(c) Silhouette (higher-better)



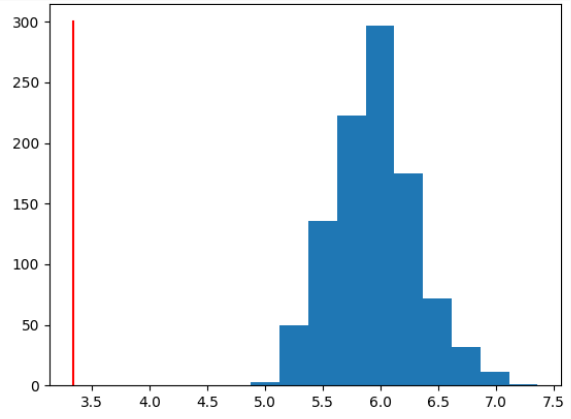
(d) Adjusted Rand (higher-better)

Distribution of quality scores for 1000 random labelings (blue histogram) vs. ground-truth scores (red line).

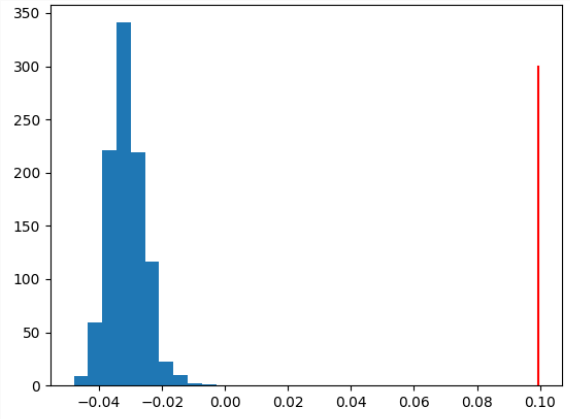
Clustering permutation plots: RoBERTa (d=42)



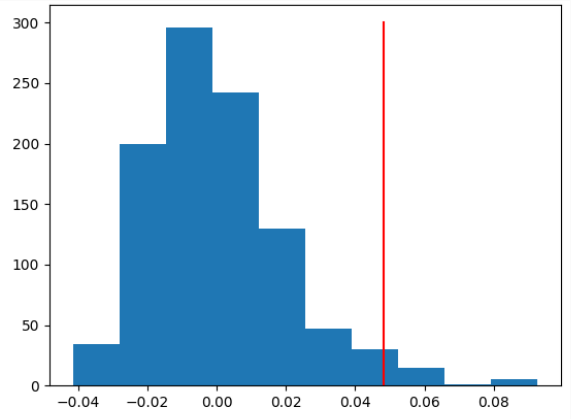
(a) Calinski-Harabasz (higher-better)



(b) Davies-Bouldin (lower-better)



(c) Silhouette (higher-better)



(d) Adjusted Rand (higher-better)

Distribution of quality scores for 1000 random labelings (blue histogram) vs. ground-truth scores (red line).