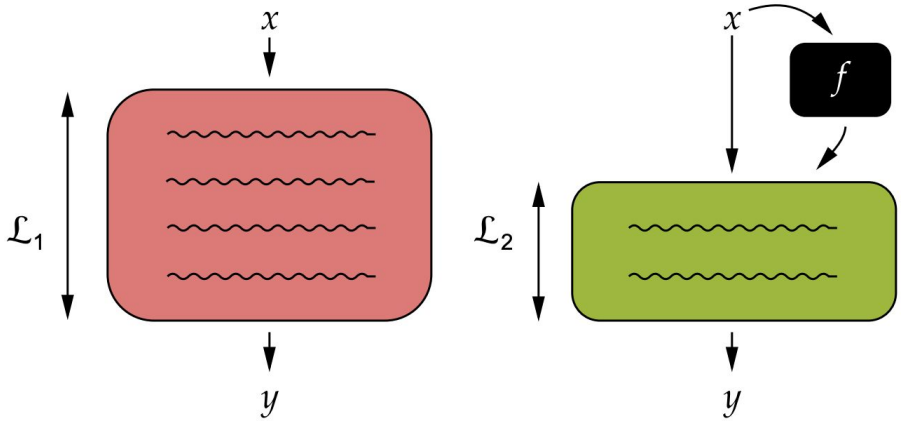
Rissanen Data Analysis:
Examining Dataset Characteristics via Description Length



Ethan Perez, Douwe Kiela, Kyunghyun Cho

Problem

What capabilities help to achieve a good model of the data?

Problem

What capabilities help to achieve a good model of the data?

1. For ImageNet, does it help to model the background?
2. For parole classification data, does it help to know a person's race?
3. For question-answering data, does it help to ask/answer subquestions?

Setup

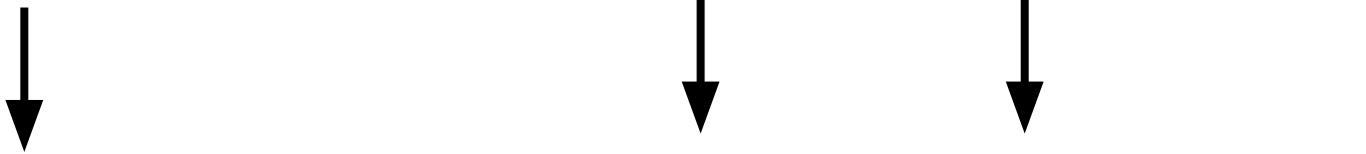
What capabilities help to achieve a good model of the data?

$f(x)$

$maps\ x \rightarrow y \quad \forall\ (x, y)$

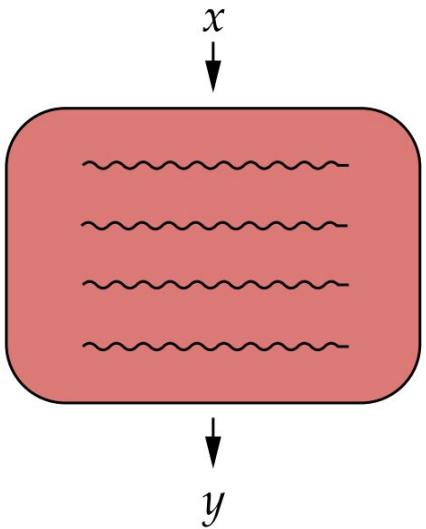
Setup

What capabilities help to achieve a good model of the data?



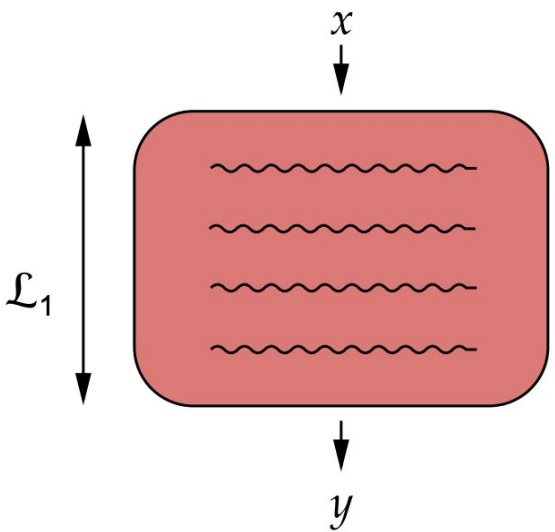
Approach

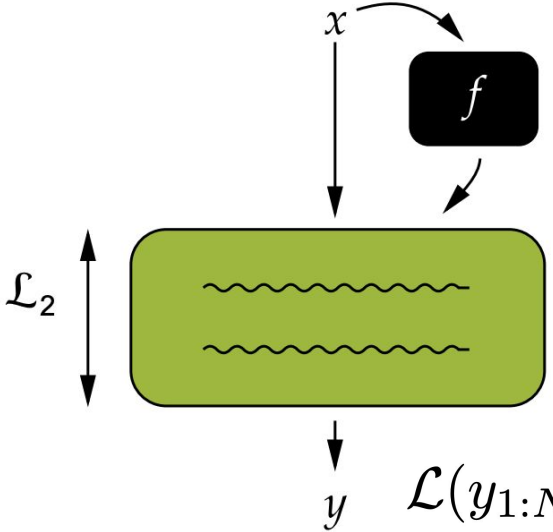
Let's view each label ***y*** as generated by a **program** executed over the input ***x***.



Approach

We say ***f(x)*** is helpful if it reduces the length of the shortest program needed to generate the data labels.

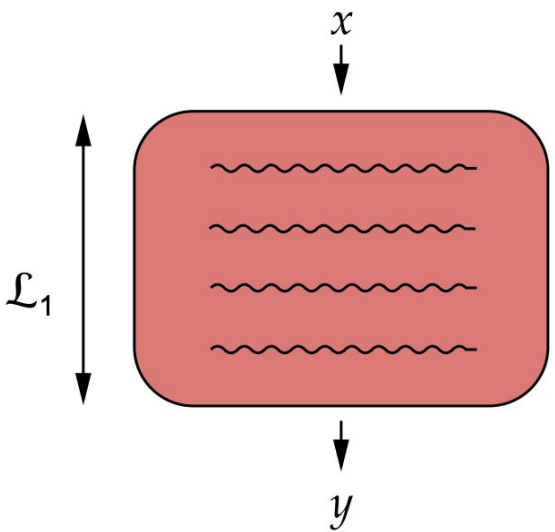


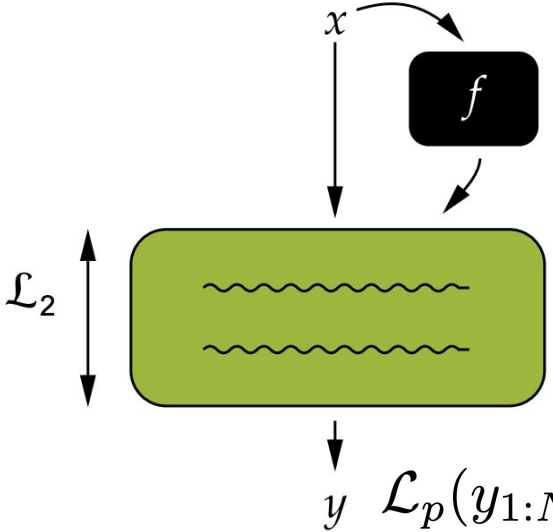


$$\mathcal{L}(y_{1:N}|x_{1:N}, f) < \mathcal{L}(y_{1:N}|x_{1:N})$$

Approach

Shortest program length is uncomputable, so we estimate the labels' *Minimum Description Length* as a proxy.





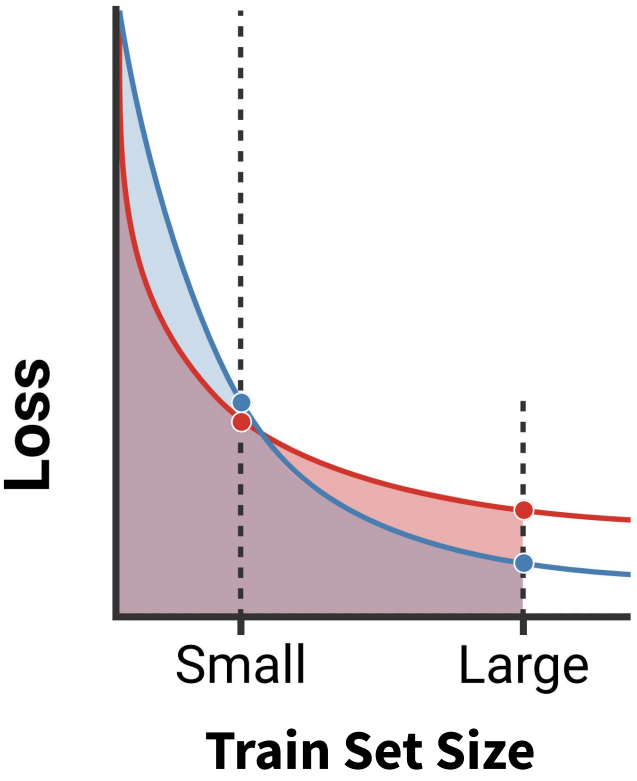
$$\mathcal{L}_p(y_{1:N}|x_{1:N}, f) < \mathcal{L}_p(y_{1:N}|x_{1:N})$$

Minimum Description Length

Measured with *"online coding"*:

- Area under the "online" learning curve when training $\boldsymbol{\theta}_{\boldsymbol{n}}$ on the first ***n-1*** examples:

$$\mathcal{L}_p(y_{1:N}|x_{1:N}) = \sum_{n=1}^{N} -\log_2 p_{\theta_n}(y_n|x_n)$$



Minimum Description Length

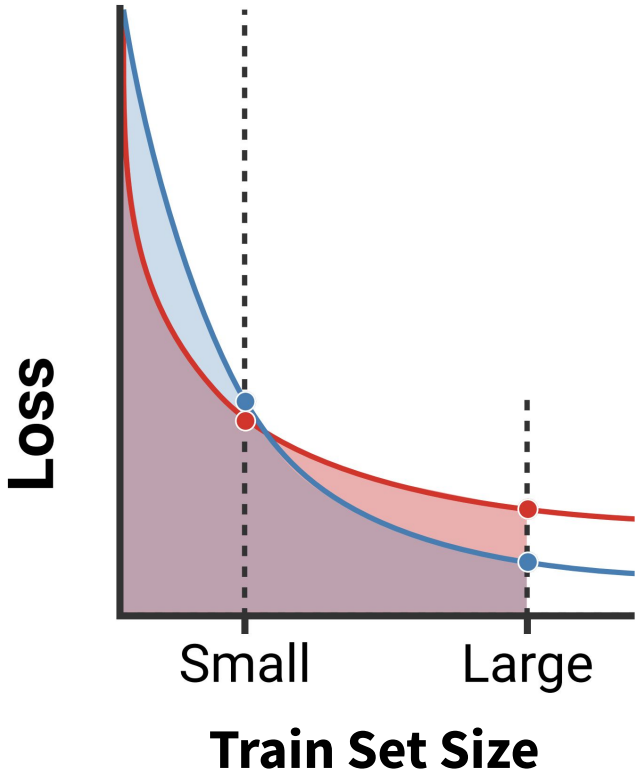
Measured with *"online coding"*:

- Area under the "online" learning curve when training $\boldsymbol{\theta}_{\boldsymbol{n}}$ on the first ***n-1*** examples:

$$\mathcal{L}_p(y_{1:N}|x_{1:N}) = \sum_{n=1}^{N} -\log_2 p_{\theta_n}(y_n|x_n)$$

$$\mathcal{L}_p(y_{1:N}|x_{1:N}, f) < \mathcal{L}_p(y_{1:N}|x_{1:N})$$

Rissanen Data Analysis (RDA)



Are subquestions useful for HotpotQA?

Example question decompositions from *Perez et al. 2020*

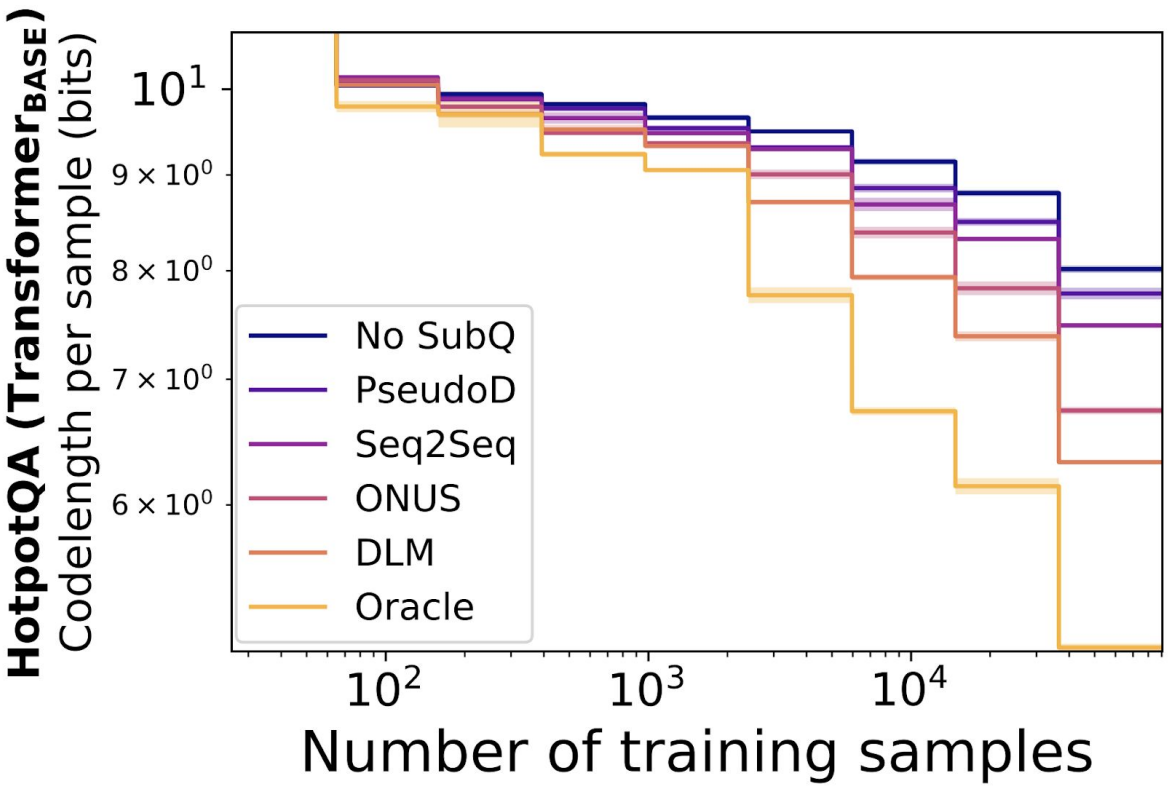
Q1: Who is older, Annie Morton or Terry Richardson?
  SQ$_1$: Who is Annie Morton?
  ∟ Annie Morton (born October 8, 1970) is an American model born in Pennsylvania.
  SQ$_2$: When was Terry Richardson born?
  ∟ Kenton Terry Richardson (born 26 July 1999) is an English professional footballer who plays as a defender for League Two side Hartlepool United.
Â: Annie Morton

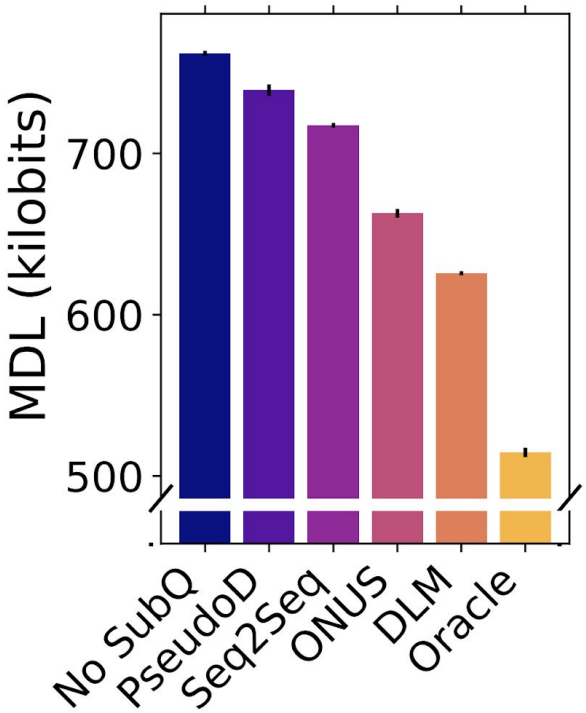
Are subquestions useful for HotpotQA?

Compare MDL of answers when just giving a QA model:

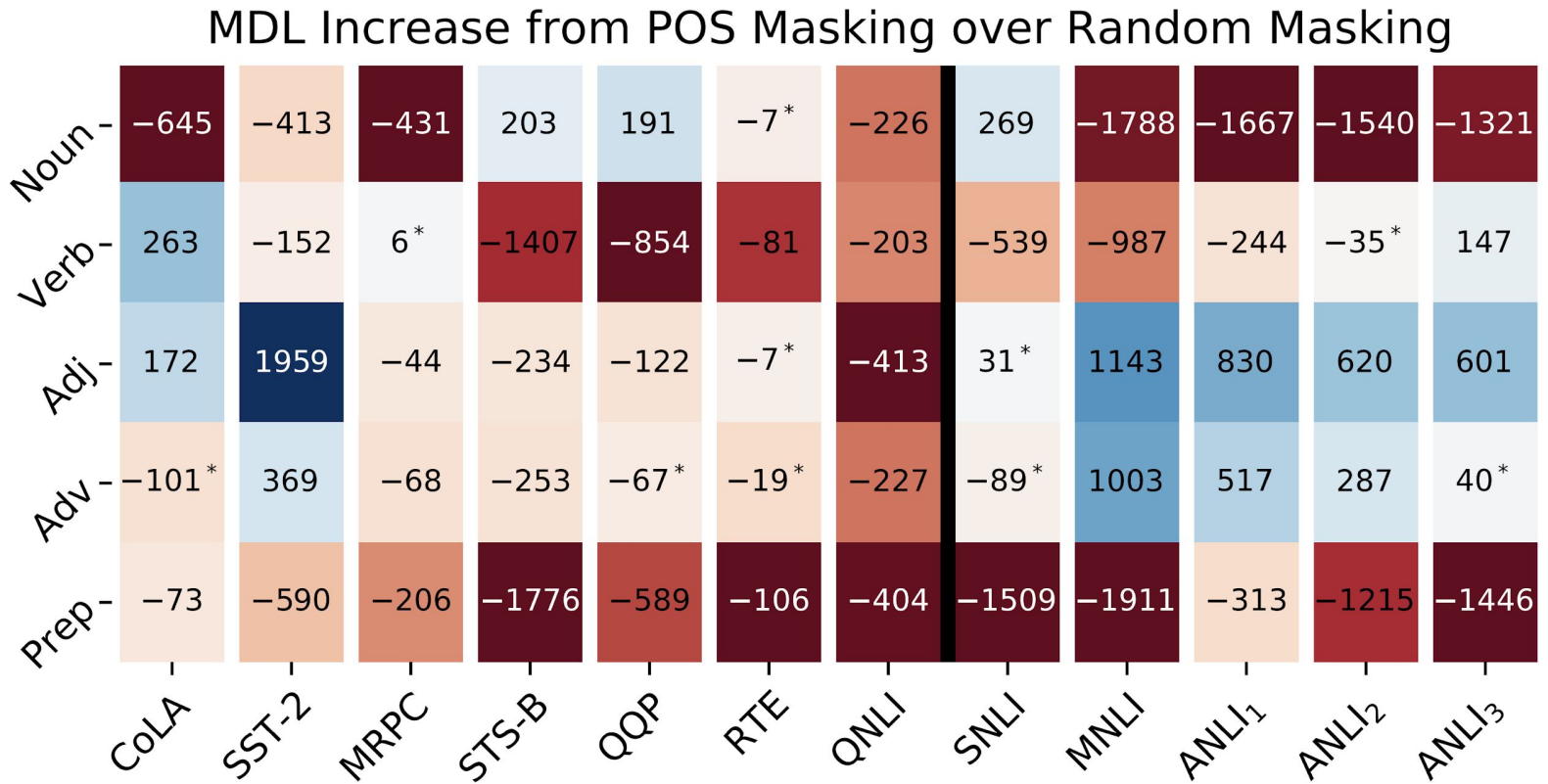
1. **Context**, **Q**
2. **Context**, **Q**, + **indicators** about which paragraph contains the answers to subquestions

Are subquestions useful for HotpotQA?

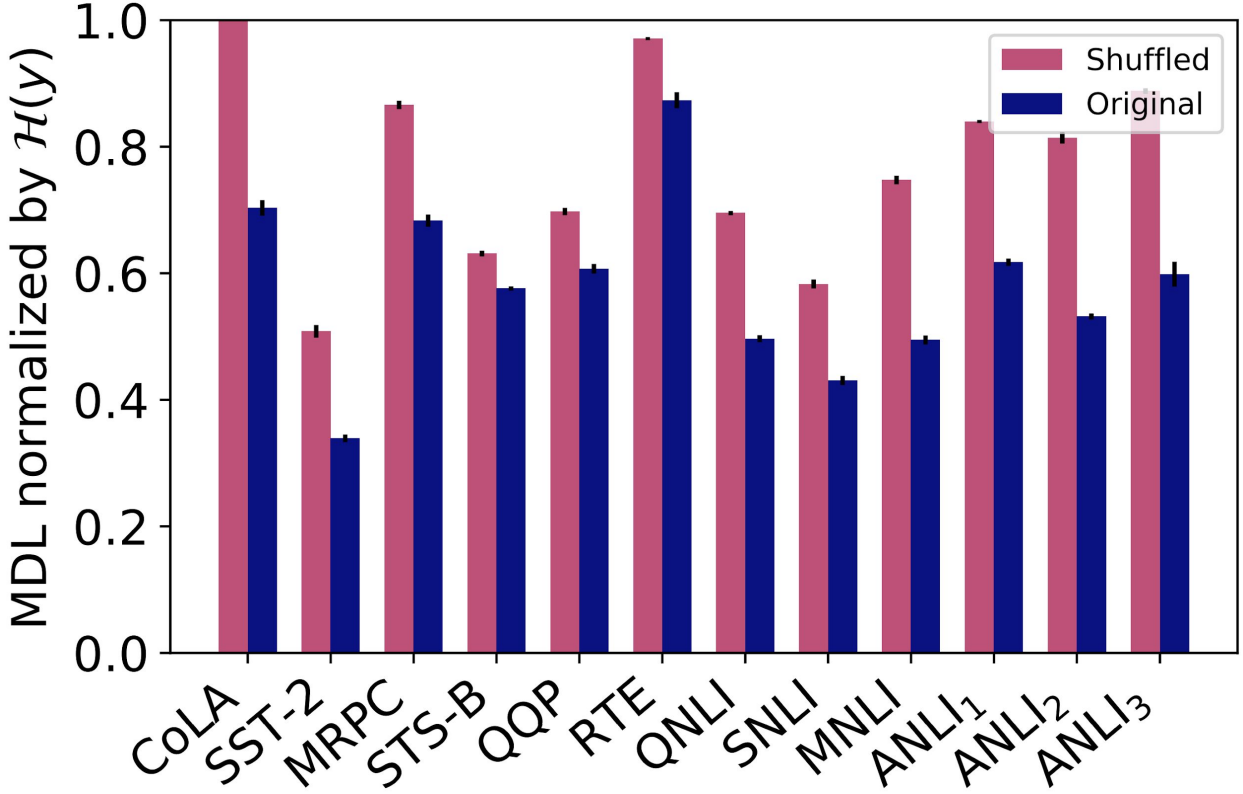




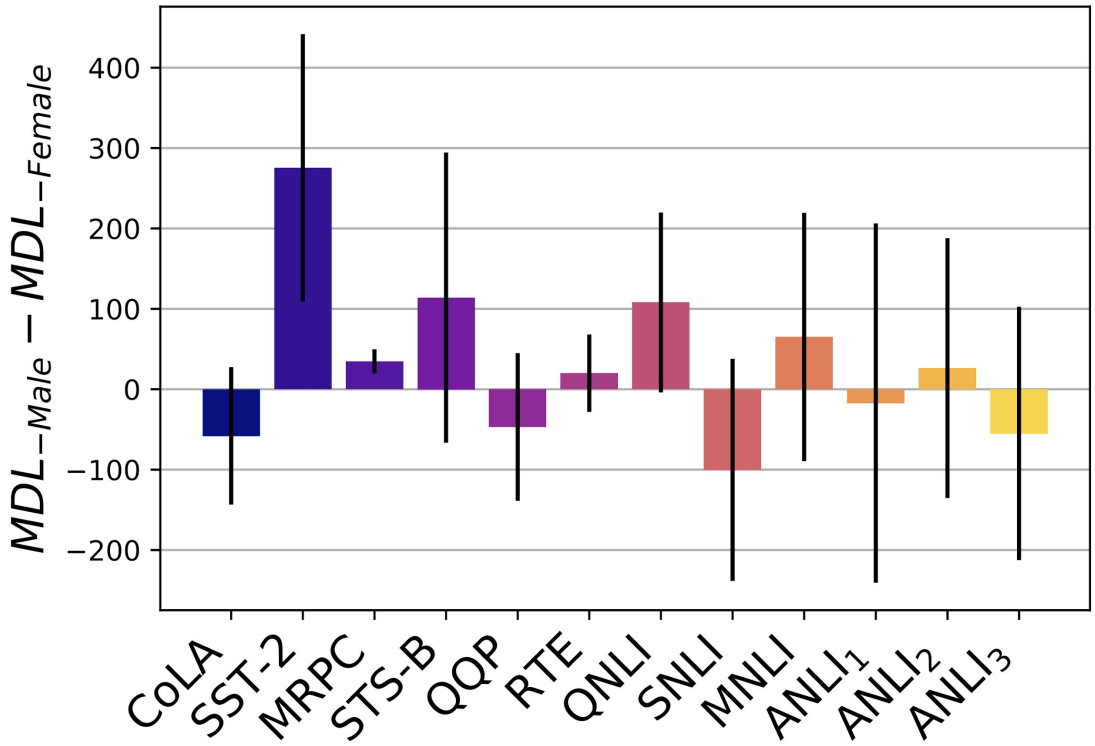
What POS are useful for NLP tasks?



Is word order useful?



Are male-gendered words more useful than female-gendered words?



Conclusion

RDA lets us:

- Evaluate what capabilities help/hurt on different datasets
- Analyze our datasets for unintended biases and artifacts
- Predicting Generalization?
  - See *True Few-Shot Learning with Language Models*, Perez, Kiela, Cho 2021