

# Reproducing the Argument Quality Prediction of Project Debater

## Motivation

- Predicting argument quality is crucial for many applications
- Model from IBM (Project Debater) not longer accessible
- Goal:** Reproduce model and make it available again

## Reproducing the IBM model

- Use a pre-trained BERT base model (uncased)
- Add a linear layer and a sigmoid activation function
- Additionally add a dropout layer (with  $p = 0.1$ )
- Finetune model on original IBM dataset

## Evaluation on Original Testset

Rank correlation with ground truth (Pearson's  $r$ , Spearman's  $\rho$ )

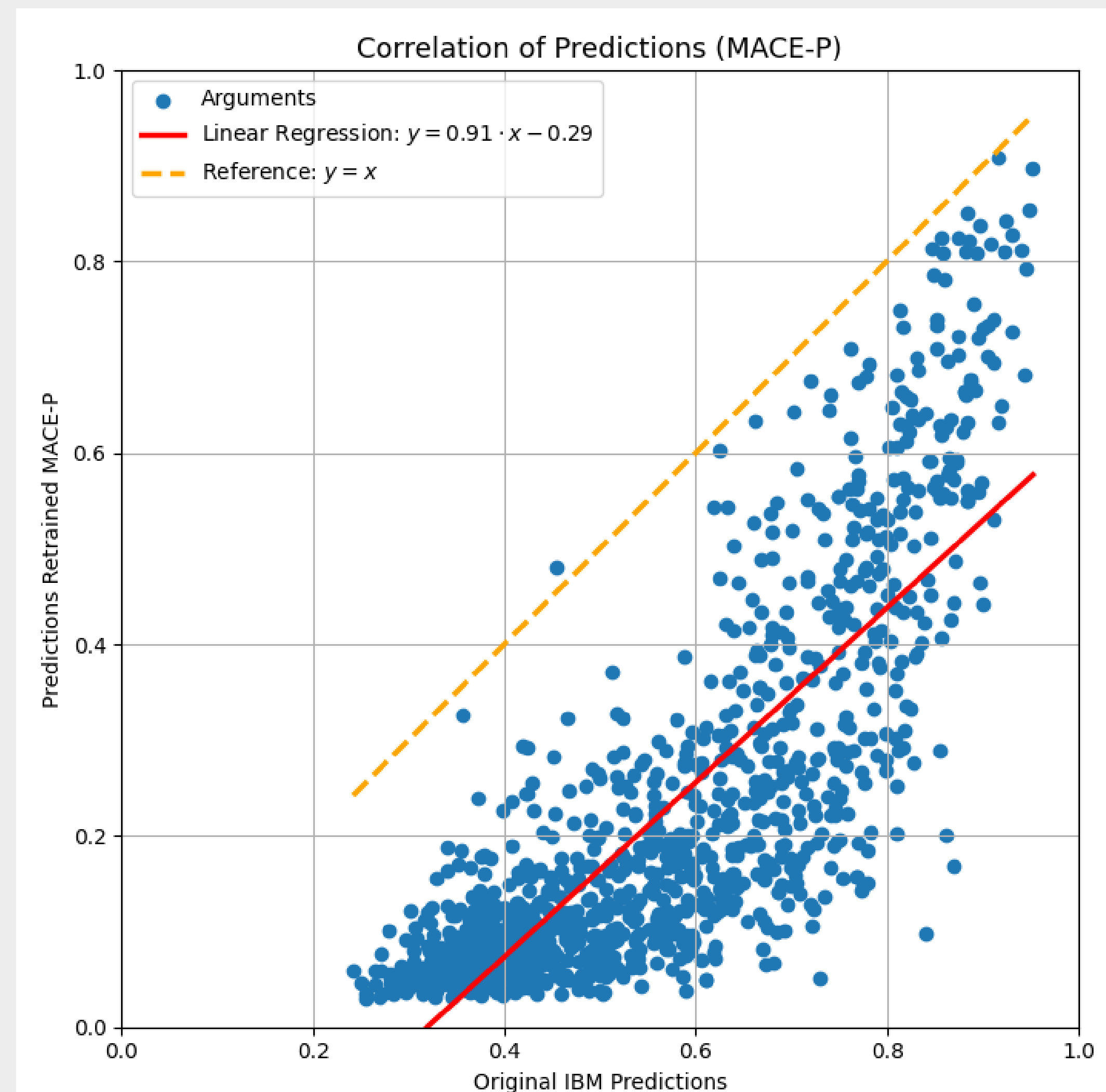
Model	MACE-P		WA	
	$r$	$\rho$	$r$	$\rho$
Original	0.53	0.52	0.52	0.48
Reproduction attempt				
1 epoch	0.537	0.523	0.532	0.482
2 epochs	0.533	0.522	0.536	0.487
3 epochs	0.485	0.480	0.494	0.441

## Evaluation on External Data from Args.me

- RMSE between scores of IBM and retrained models
- Number of lower ( $\searrow$ ) and higher ( $\nearrow$ ) predicted scores
- Average distance to IBM score for lower and higher scores
- Correlation between scores of IBM and reproduced models

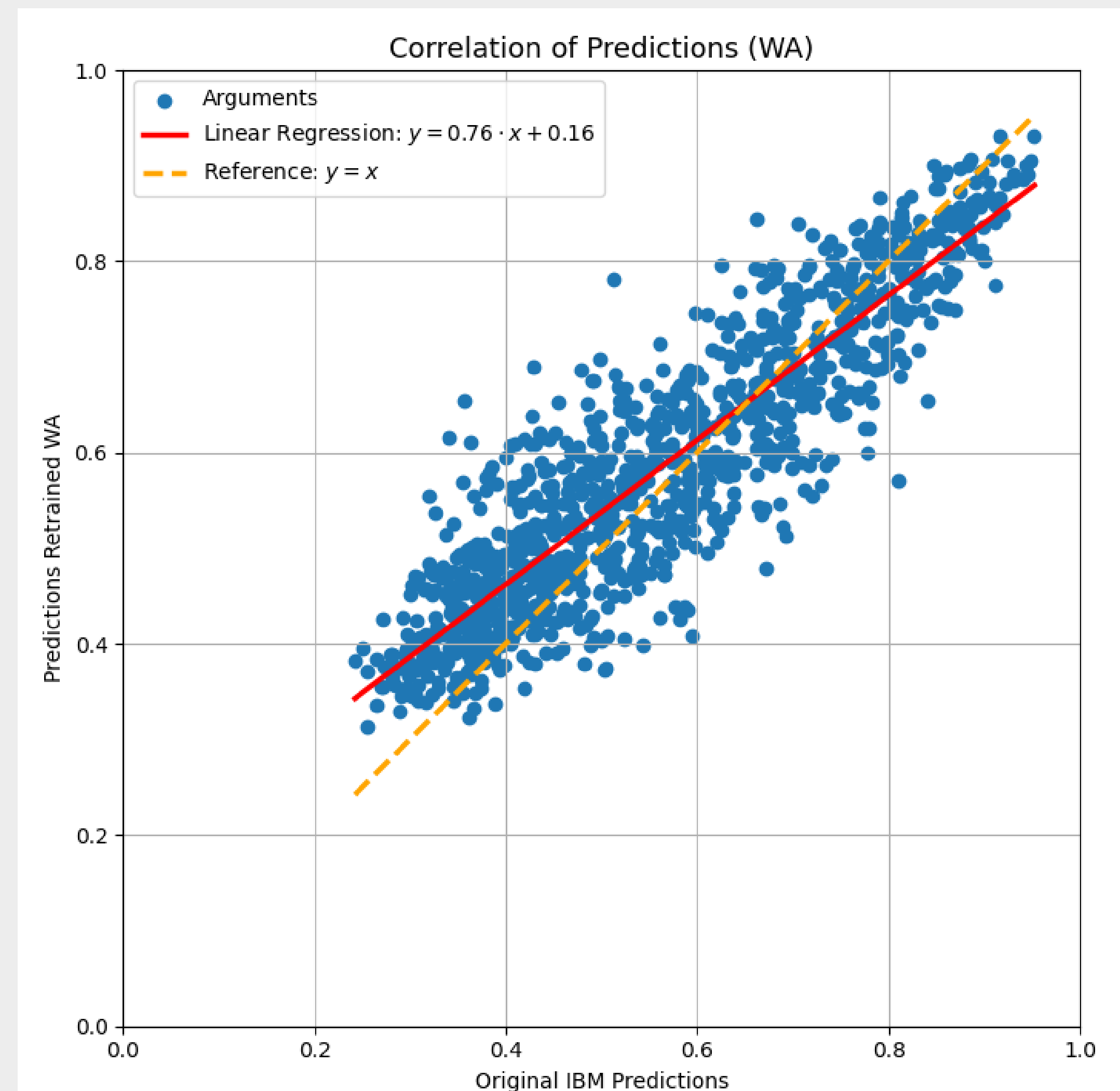
### MACE-P

RMSE	# $\searrow$	$\searrow$	# $\nearrow$	$\nearrow$	$r$	$\rho$
0.359	1132	0.34	1	0.03	0.816	0.823



### WA

RMSE	# $\searrow$	$\searrow$	# $\nearrow$	$\nearrow$	$r$	$\rho$
0.080	428	0.05	705	0.07	0.901	0.889



## Qualitative Evaluation

Sentence (from args.me)	IBM	MACE-P	WA
Now as for the definition.	0.39	0.14	0.48
Evidence show the DP is more expensive.	0.45	0.08	0.48
But does this make it right to kill them back?	0.43	0.06	0.47
So there are undoubtedly instances in the past where we have executed an innocent man but did not know so, and still do not know.	0.82	0.29	0.69
The military is in no obligation to let women into the frontlines just because they hold 95% of the Armies positions [...].	0.52	0.18	0.65

## Findings

- Similar effectiveness on original data (two training epochs)
- Stronger correlation on external texts for WA model
- Not clear whether WA or IBM model generalizes better

## Resources

- <https://github.com/webis-de/argmining25-reproducing-ibm-arg-quality-api>
- [webis.de/publications.html#zelch\\_2025b](https://webis.de/publications.html#zelch_2025b)

