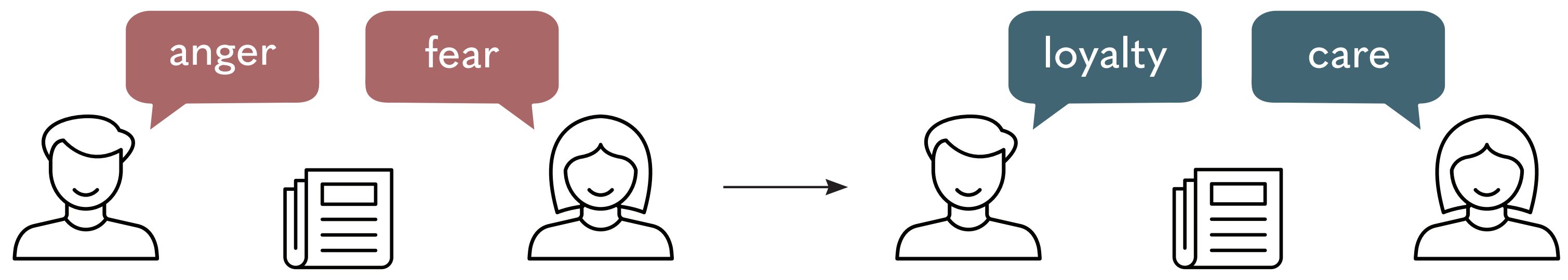
Predicting Value Interpretations from SEAT Annotations

Value Interpretations

Can we predict a person's interpretation of values in text from their judgment of other subjective dimensions (Sentiment, Emotion, Argument, Topic)?



Dataset

50 justifications provided by citizens in an energy transition survey, annotated by 5 annotators with SEAT dimensions and values, with different levels of annotator agreement:

Sentiment	Emotion	Argument	Topic	Values
0.17	0.00365	0.2447	0.514	0.0144

Method

We prompt Llama-3.1-8B-Instruct zero-shot (providing the list of values to choose from). For each annotator, 20 + 1 variants.

	Sentiment	Emotion	Argument	Topic	All
One-shot	OS-S	OS-E	OS-A	OS-T	OS-all
Few-shot (5)	FS-5-S	FS-5-E	FS-5-A	FS-5-T	FS-5-all
Few-shot (10)	FS-10-S	FS-10-E	FS-10-A	FS-10-T	FS-10-all
Few-shot (15)	FS-15-S	FS-15-E	FS-15-A	FS-15-T	FS-15-all

Zero-shot (ZS) baseline:

```
> What values are expressed in this justification?
```

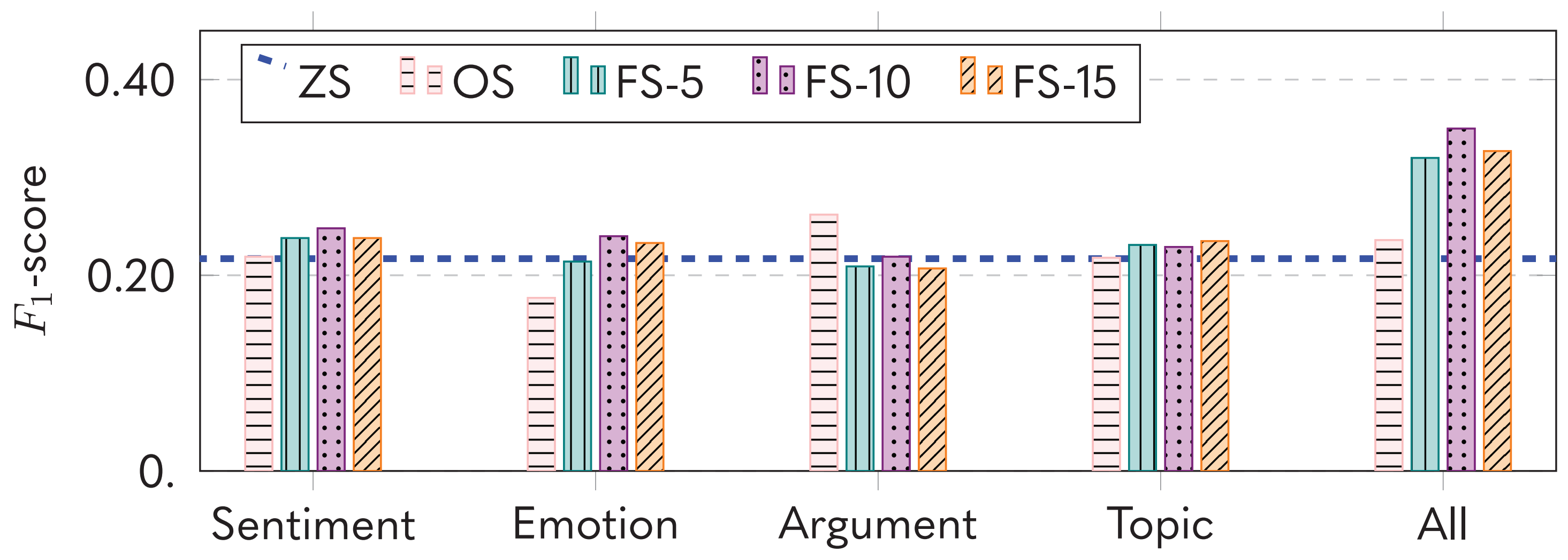
One-shot (OS):

```
> What values are expressed in this justification,
given how this person annotated this justification
with this S/E/A/T dimension?
```

Few-shot (FS):

```
> What values are expressed in this justification,
given how this person annotated this and other K
justifications with this S/E/A/T dimension?
```

Providing all dimensions helps



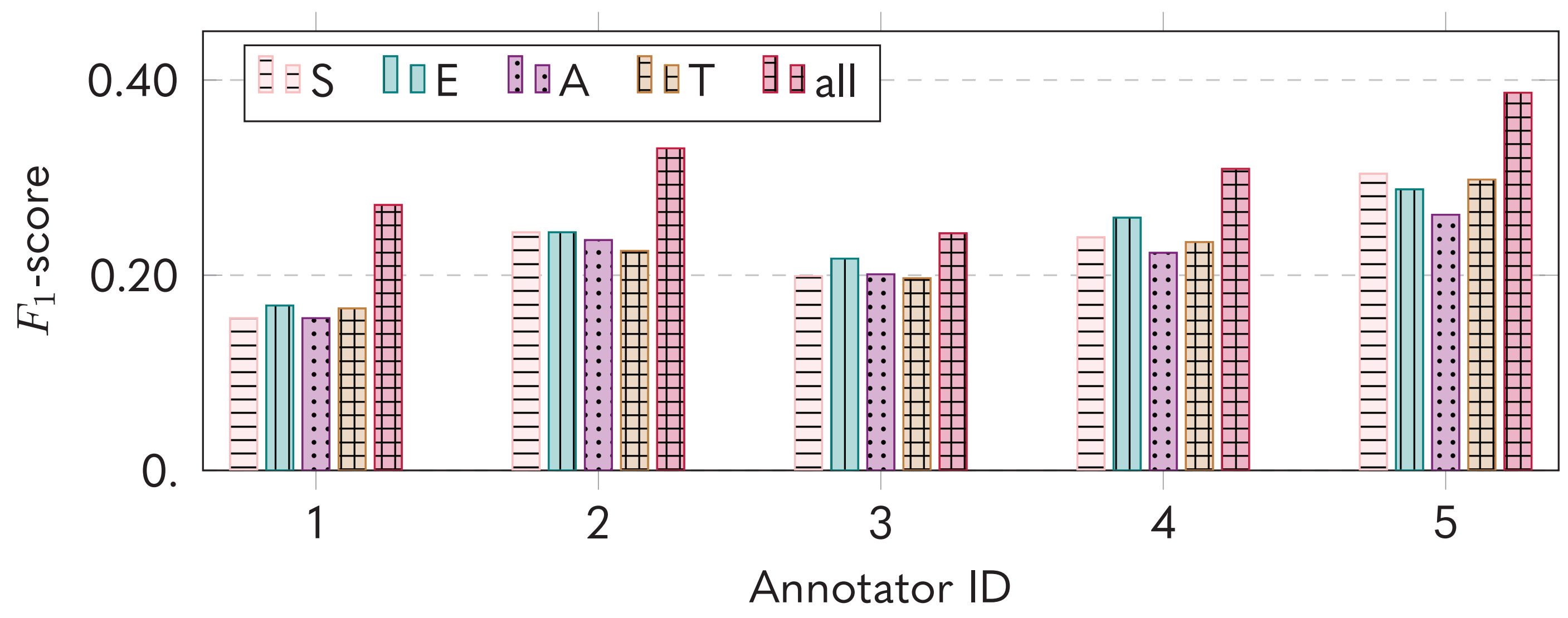
Takeaways

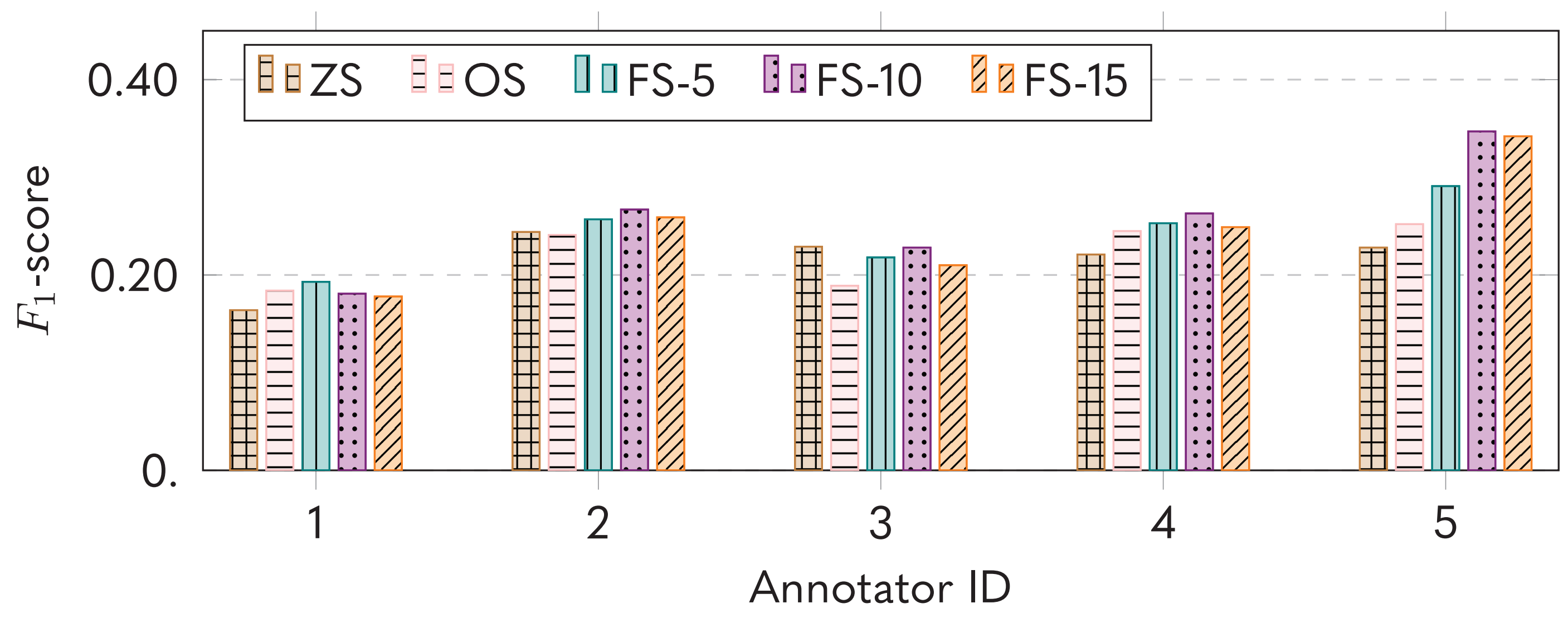
Providing a few in-context examples with all SEAT dimensions works.

The performance is far from perfect.

Differences across individuals

Consistent trends, but different results.





"Taking a SEAT: Predicting Value Interpretations from Sentiment, Emotion, Argument, and Topic Annotations". A.N. Dobrinoiu, A.C. Marcu, A. Homayounirad, L. Cavalcante Siebert, E. Liscio. VALE @ ECAI'25.