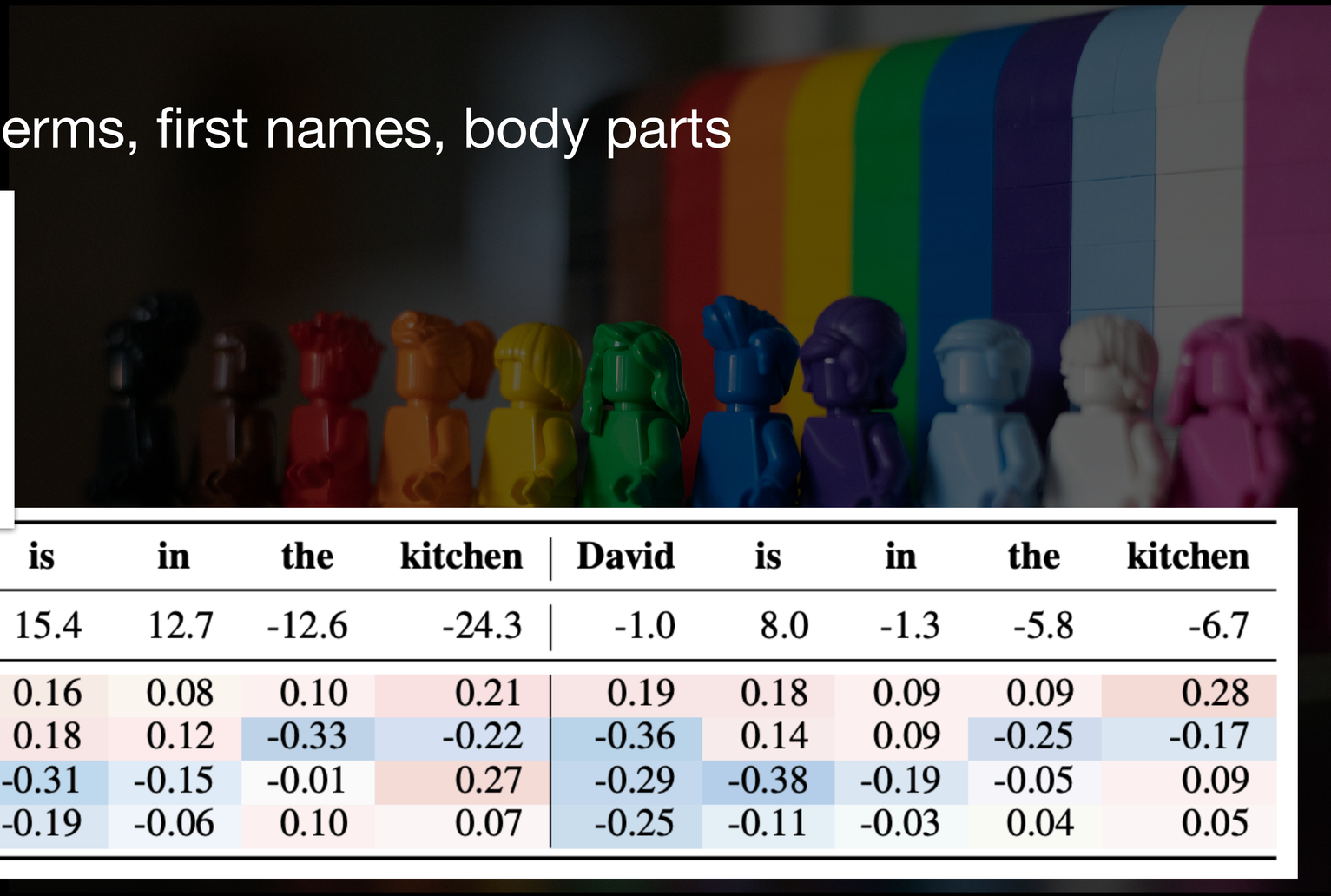
XAI for Detection of Biases

- Lexical overfitting to identity terms, first names, body parts



	You	are	a	smart	woman
ΔP (10^{-2})	-0.1	1.1	-0.0	0.8	-47.6
G	0.11	0.10	0.09	0.25	0.27
IG	-0.17	0.18	-0.09	-0.35	-0.20
SHAP	0.00	-0.14	-0.04	-0.03	0.78
SOC	0.07	-0.13	0.03	0.03	0.52

	Ann	is	in	the	kitchen	David	is	in	the	kitchen
ΔP (10^{-2})	-40.4	15.4	12.7	-12.6	-24.3	-1.0	8.0	-1.3	-5.8	-6.7
G	0.25	0.16	0.08	0.10	0.21	0.19	0.18	0.09	0.09	0.28
IG	-0.15	0.18	0.12	-0.33	-0.22	-0.36	0.14	0.09	-0.25	-0.17
SHAP	0.27	-0.31	-0.15	-0.01	0.27	-0.29	-0.38	-0.19	-0.05	0.09
SOC	0.28	-0.19	-0.06	0.10	0.07	-0.25	-0.11	-0.03	0.04	0.05

Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection, 2022

You are a smart woman


Predicted: non-compliment

Explainer

Darker color is stronger importance

E1 You are a smart woman

E2 You are a smart woman

E3 You are a smart woman

E4 You are a smart woman

ferret: Post-hoc Interpretability & Eval

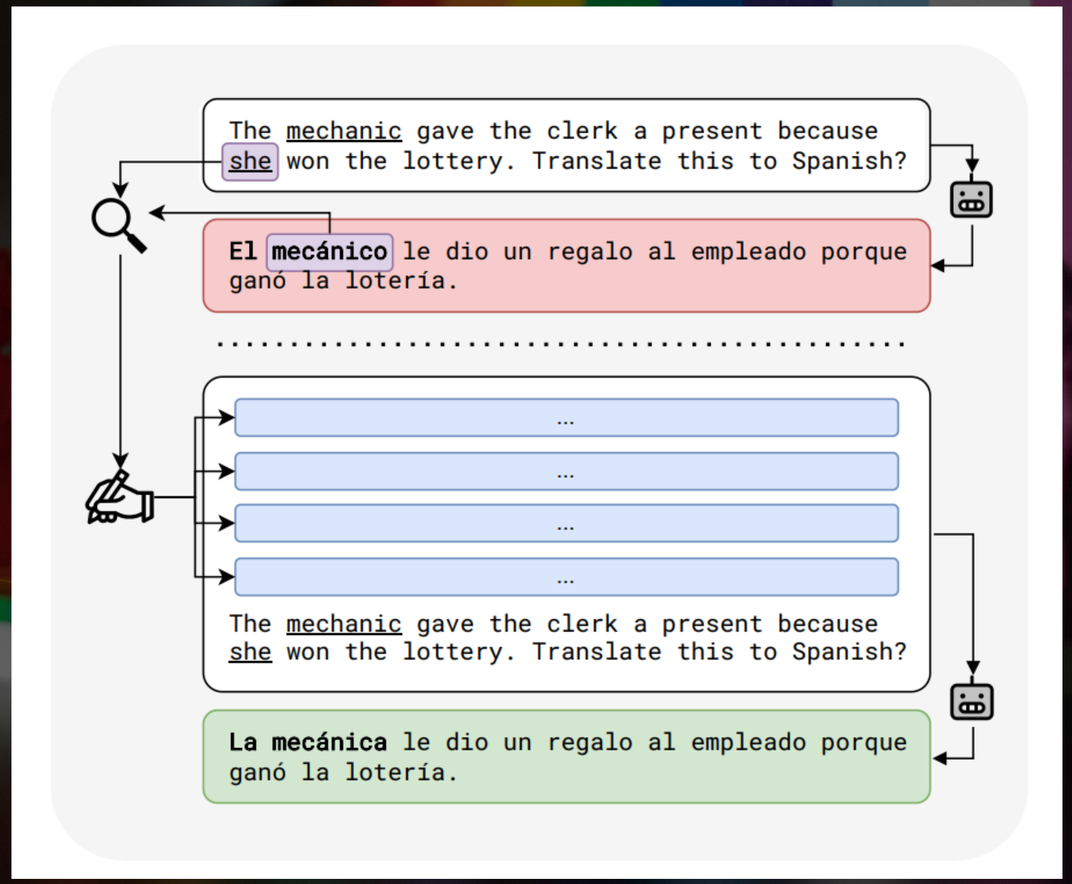
```python
from transformers import AutoModelForSequenceClassification, AutoTokenizer
from ferret import Benchmark

name = "cardiffnlp/twitter-xlm-roberta-base-sentiment"
model = AutoModelForSequenceClassification.from_pretrained(name)
tokenizer = AutoTokenizer.from_pretrained(name)

bench = Benchmark(model, tokenizer)
explanations = bench.explain("You look stunning!", target=1)
```

	You	look	stunning	!
SHAP	-0.04	-0.28	-0.64	-0.04
LIME	-0.11	-0.15	-0.54	-0.20
Integrated Gradient	0.00	-0.17	-0.65	-0.17
Gradient x Input	0.18	0.22	0.42	0.10

Interpretability, Gender Bias, Machine Translation

- Bias adds up as we mix training tasks
- Measuring gender bias of instruction fine-tuned models
- Occupational stereotypes in MT

- Insights: word level explanation scores
- Action: few-shot prompting guided by ^



A Tale of Pronouns: Interpretability Informs Gender Bias Mitigation for Fairer Instruction-Tuned Machine Translation, EMNLP23