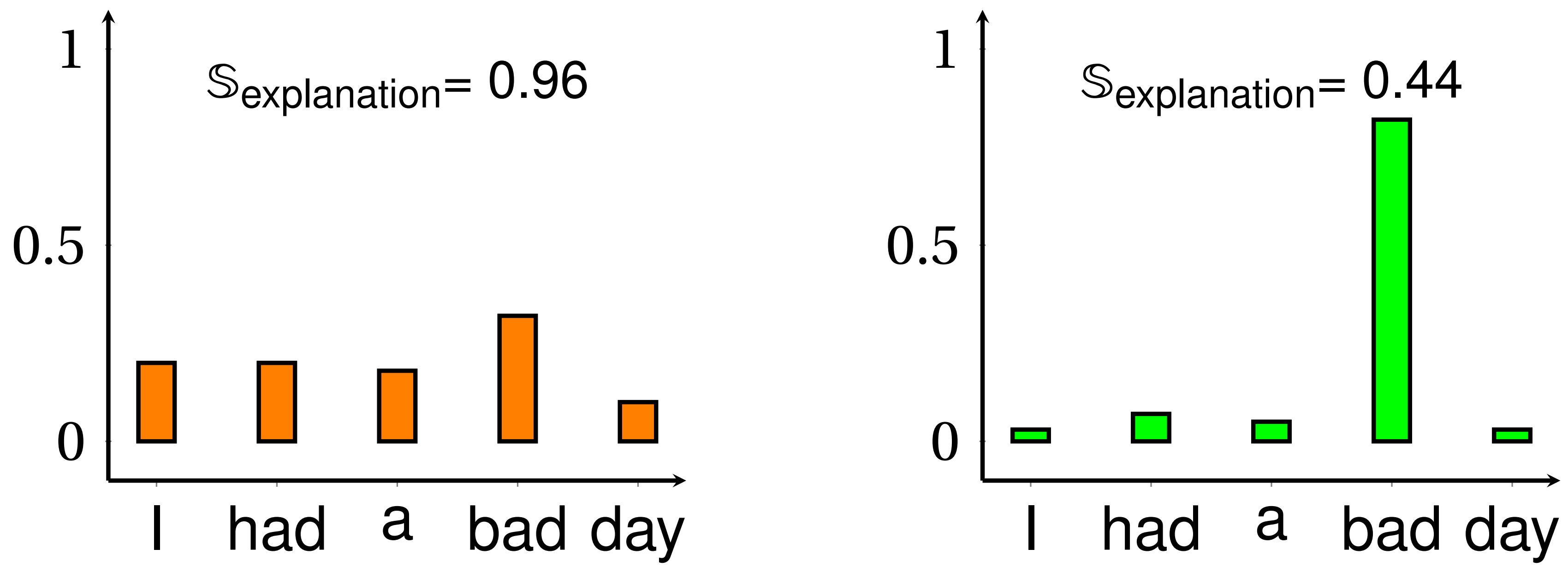
Simplicity is not as *simple* as you think

Aman SINHA[1,2], Timothee MICKUS[3], Xavier COUBEZ[2], Marianne CLAUSEL[1], Mathieu CONSTANT[1]

What is a *simple* Explanation?

Based on previous work (Bhatt et al., 2020),

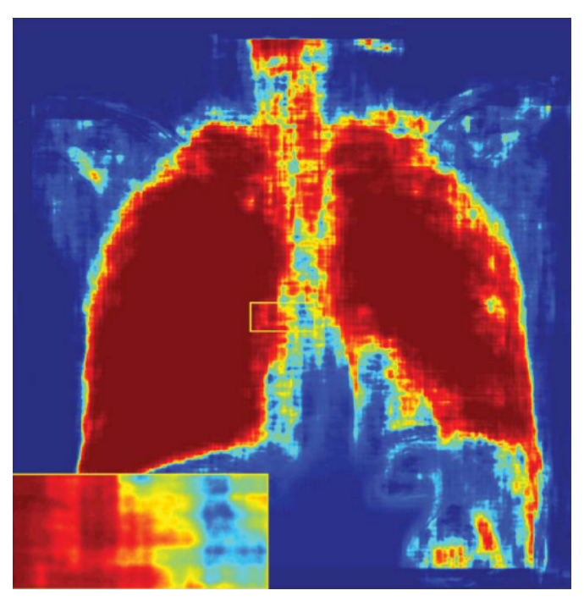
$$\mathbb{S}_{\text{explanation}} = -\sum \hat{w}_i \log_2 \hat{w}_i$$



Bhatt et al., 2020 *Evaluating and aggregating feature-based model explanations.* arXiv preprint arXiv:2005.00631.

Motivation

There are expected human-centric properties from an explanation such as:
(i) the semantic alignment between with user's mental model (aka. rationales)
(ii) the presentation in the explanation (also refereed to as understandability, cf. Chen et al. (2022); Moreno-Sánchez (2023))



- ► A tumor should ideally be identified on the basis of the limited set of pixels corresponding to said tumor and its accompanying telltale signs.
- ► An equally accurate decision support system that would highlight the entirety of the image would be found less useful.

Disagreement in human annotations (Sundararajan & Najmi, 2020; Atanasova, 2024) is another factor that tends to deviate the model from expected behavior (Mickus et al., 2025) which can potentially be reflected in explanations.

Chen et al., 2022. *What makes a good explanation?: A harmonized view of properties of explanations.* arXiv preprint arXiv:2211.05667
Mickus et al., 2025. *Your Model is Overconfident, and Other Lies We Tell Ourselves.* In Proceedings of the 63rd Annual Meeting of the ACL (Volume 2: Long Papers).

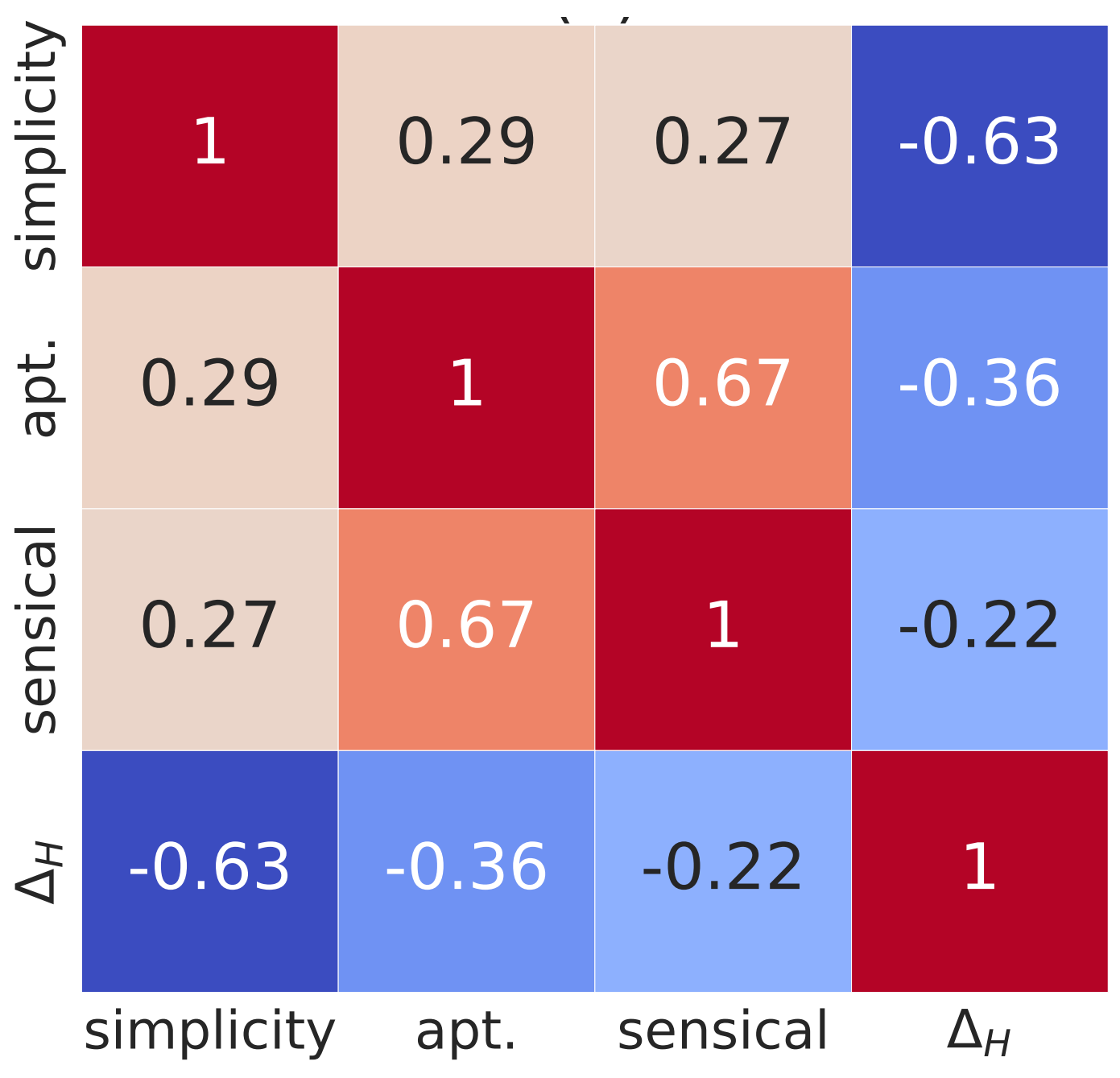
Evidence

I. Human have a moderate preference for simplicity.

Setup : For SST2 dataset (Socher et al., 2013) with 24 BERT models' (Turc et al., 2019) generated explanations, we ask 3 annotators to select from the two paired explanations fit best for each the following criteria: (a) simplicity (b) appropriateness (c) sensicality

Criteria	Description
simplicity	which explanation assigns weights to fewer words (concise)?
appropriate (apt.)	which puts greater emphasis on words you would have paid attention?
sensicality	Which is easy to make sense of (i.e reconstruct a reasoning)?
Δ_H	$\lvert\mathtt{max}(\mathbb{S}_{\text{expl.}})$ - $\mathtt{min}(\mathbb{S}_{\text{expl.}})\rvert$

Table 1: Description of each criteria and Δ_H.



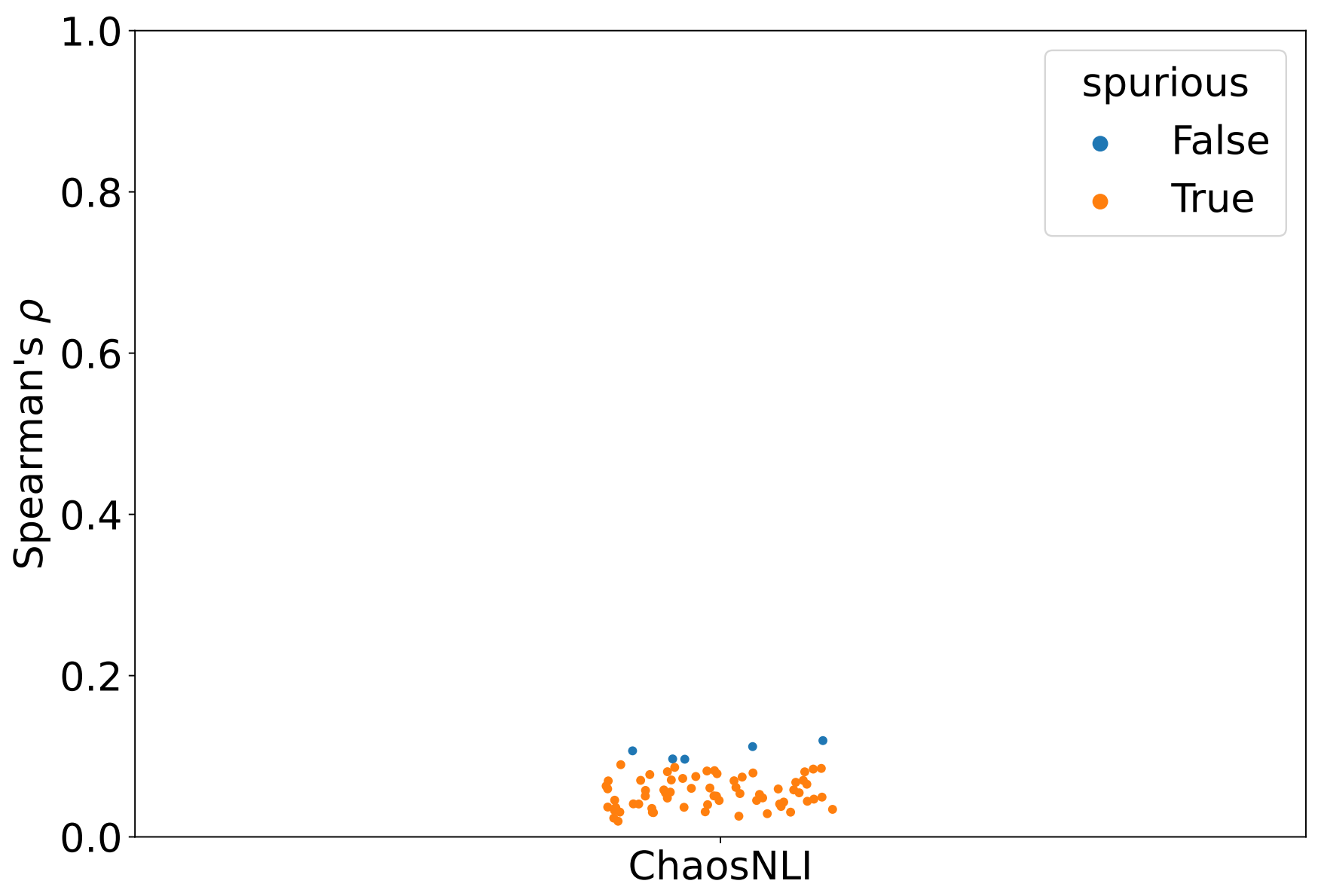
	simplicity	apt.	sensical	Δ_H
simplicity	1	0.29	0.27	-0.63
apt.	0.29	1	0.67	-0.36
sensical	0.27	0.67	1	-0.22
Δ_H	-0.63	-0.36	-0.22	1

Right confusion matrix: Preference for a simpler explanation only partially translates to a preference in terms of appropriateness and sensicality.

Socher et. al 2013. *Recursive deep models for semantic compositionality over a sentiment treebank.* In Proceedings of the 2013 conference on EMNLP, pp. 1631–1642, 2013.
Turc et. al 2019. *Well-read students learn better: On the importance of pre-training compact models.* arXiv preprint arXiv:1908.08962v2, 2019.

(B) SIMPLICITY DOES NOT NEED TO RELATE WITH HUMAN LABEL VARIATION.

Setup : For any given sample from ChaosNLI dataset (Nie et al., 2020), if there is no clear consensus on what the label should be between annotators, then any models' should yield less simple explanation.



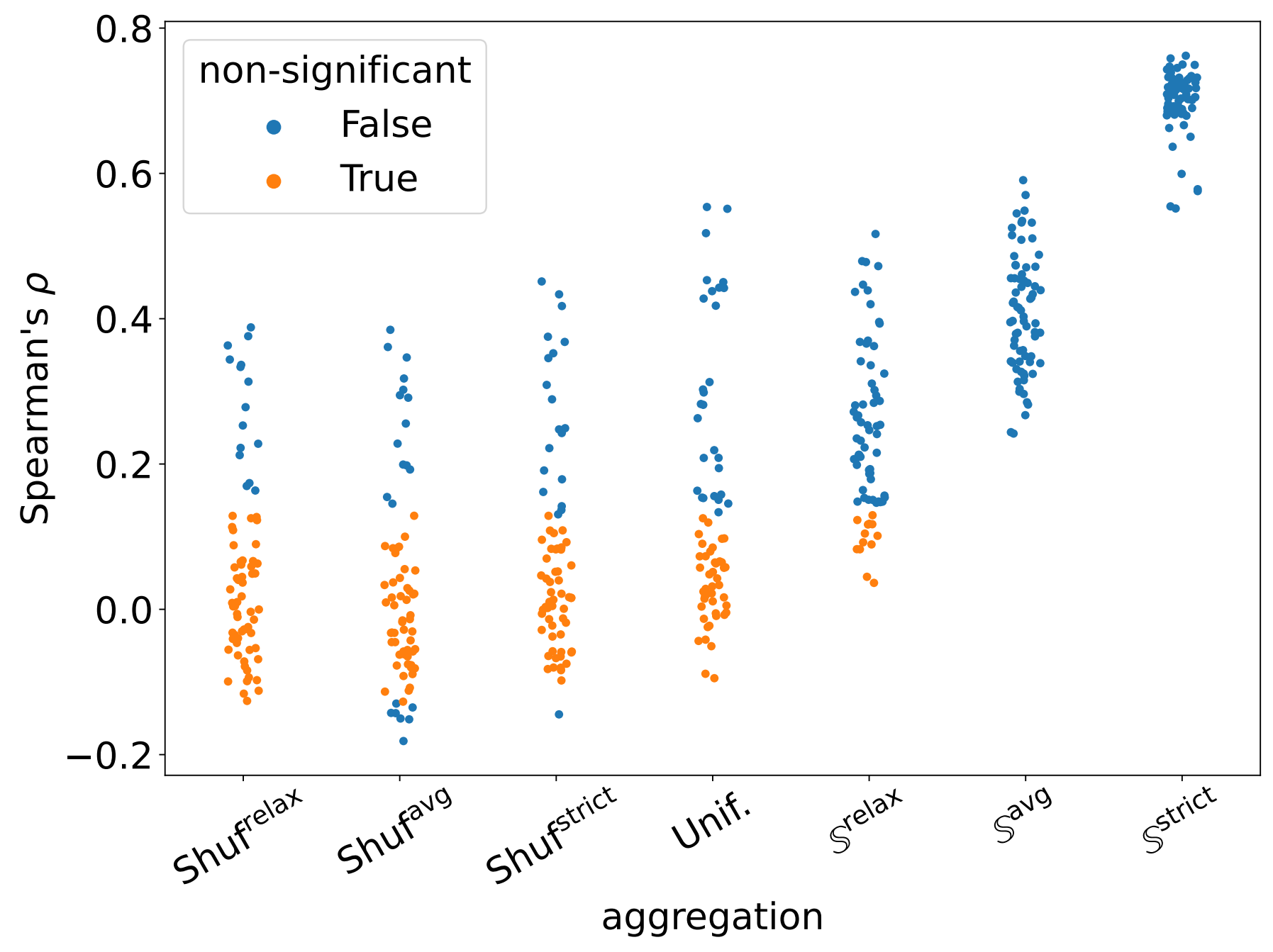
67 out of 72 of the classifiers produce spurious correlations; indicates more complex datapoints do not seem to yield more nuanced explanations. Significant correlations remain noticeably small ($0.0964 \leq \rho \leq 0.1194$).

Nie et al., 2020. *What can we learn from collective human opinions on natural language inference data?* Proceedings of the 2020 Conference on EMNLP, pp. 9131–9143.

II. Models' explanation are not always simple.

(A) SIMPLICITY TEND TO ALIGN WITH HUMAN RATIONALES.

Setup : Using HateXplain dataset (Mathew et al. 2021) and 24 BERT models' generated explanations, we compare $\mathbb{S}_{\text{explanation}}$ to the divergence between the aggregated rationale from 3 annotators versus explanations.



Aggregate	ρ is significant Min	 Mean	 Max
$\text{Shuf}^{\text{relax}}$	-0.1506	0.2316	0.3890
Shuf^{avg}	-0.1578	0.2162	0.3848
$\text{Shuf}^{\text{strict}}$	-0.1931	0.0975	0.3750
Unif.	0.1336	0.3028	0.5538
$\mathbb{S}^{\text{relax}}$	0.1295	0.2697	0.5166
$\mathbb{S}^{\text{avg}}$	0.2419	0.4061	0.5907
$\mathbb{S}^{\text{strict}}$	0.5517	0.7006	0.7619

Table 2: Corr. b/w $\mathbb{S}$ and divergence with human rationales.

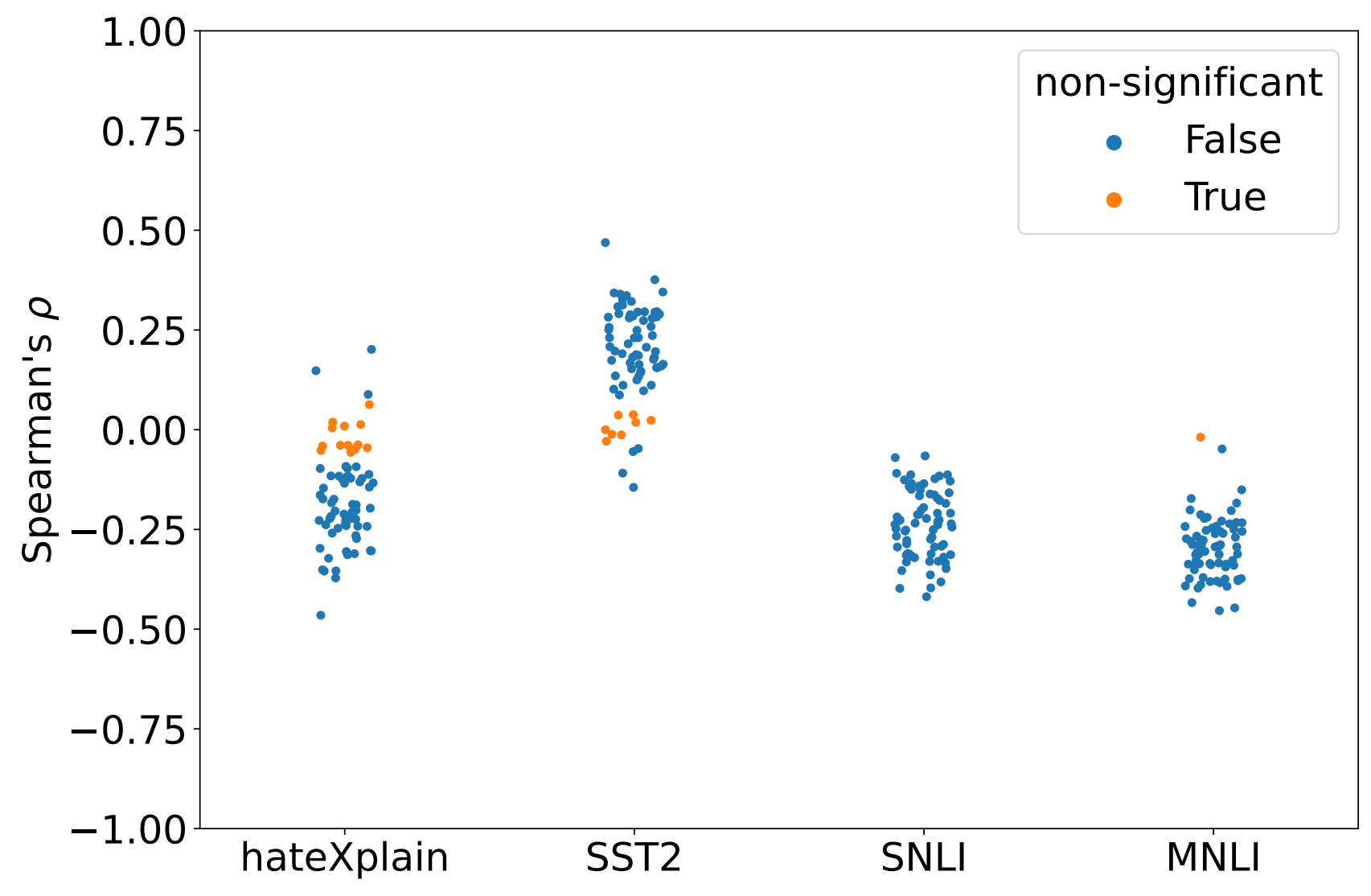
We obtain significant correlations for most of the setups we consider, underscoring that explanations match human rationales.

We also observe significant correlations with random baselines.

Mathew et al., 2021 *Hatexplain: A benchmark dataset for explainable hate speech detection.* In Proceedings of the AAAI Conference, volume 35, pp. 14867–14875, 2021.

(C) EXTRANEOUS FACTORS CAN IMPACT THE SIMPLICITY.

Setup : Using 4 datasets: HateXplain, SST2, SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), we investigate the interaction between performance of the models and simplicity of the generated explanations.



Dataset	ρ is significant Min	 Mean	 Max
SST2	-0.1445	0.2096	0.4692
HateXplain	-0.4652	-0.1955	0.2014
SNLI	-0.4186	-0.2382	-0.0657
MNLI	-0.4536	-0.2996	-0.0482

Table 3: Corr. b/w $\mathbb{S}$ and $\mathrm{Pr}(y)$.

Dataset	t statistic	Cohen's d
SST2	-67.1491	-0.7442
HateXplain	-18.7570	-0.2992
SNLI	-64.5164	-0.7148
MNLI	-51.1299	-0.5791

Table 4: T-tests b/w $\mathbb{S}_{\text{worst}}$ and $\mathbb{S}_{\text{best}}$ classifier.

Models that perform best tend to produce explanations that are more complex: i.e, models that are more successful also yield explanations that are less simple.

Bowman et al., 2015. *A large annotated corpus for learning natural language inference.* Proceedings of the 2015 Conference on EMNLP, pp. 632–642
Williams et al., 2018. *A broad coverage challenge corpus for sentence understanding through inference.* Proceedings of the 2018 Conference of the NAACL, Vol 1 (Long Papers), pp. 1112–1122

Takeaways

⚠Be cautious with explanations that are simpler!

⚠Avoid optimizing your explanations for simplicity!

AIW Workshop 2025

Affiliation
Université de Lorraine, Nancy, France[1]
ICANS Strasbourg, France[2]
University of Helsinki, Helsinki, Finland[3]

Contact
github.com/amansinha09
twitter.com/amansinha09
https://amansinha09.github.io/

Looking for POSTDOC or INDUSTRY positions in NLP / AI+Healthcare / LMs Evaluation.