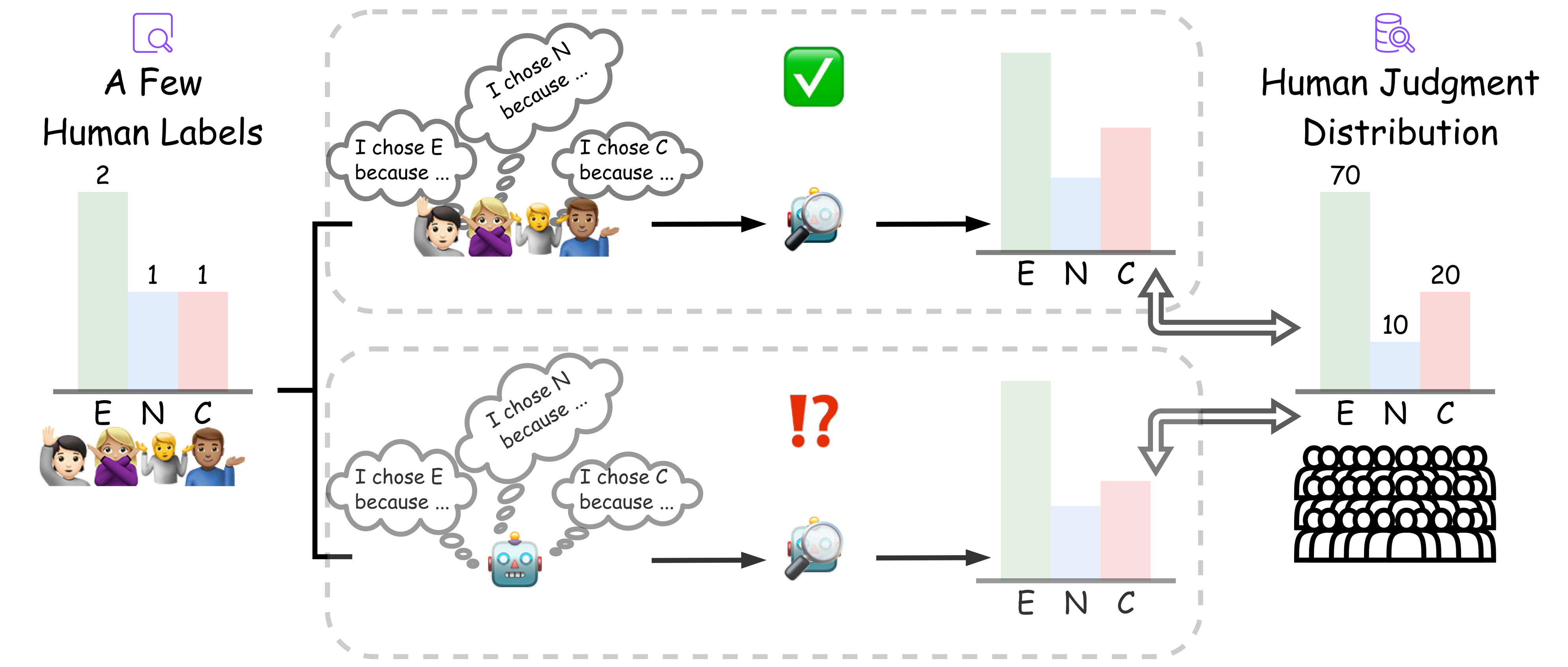
A Rose by Any Other Name: LLM-Generated Explanations Are Good Proxies for Human Explanations to Collect Label Distributions on NLI

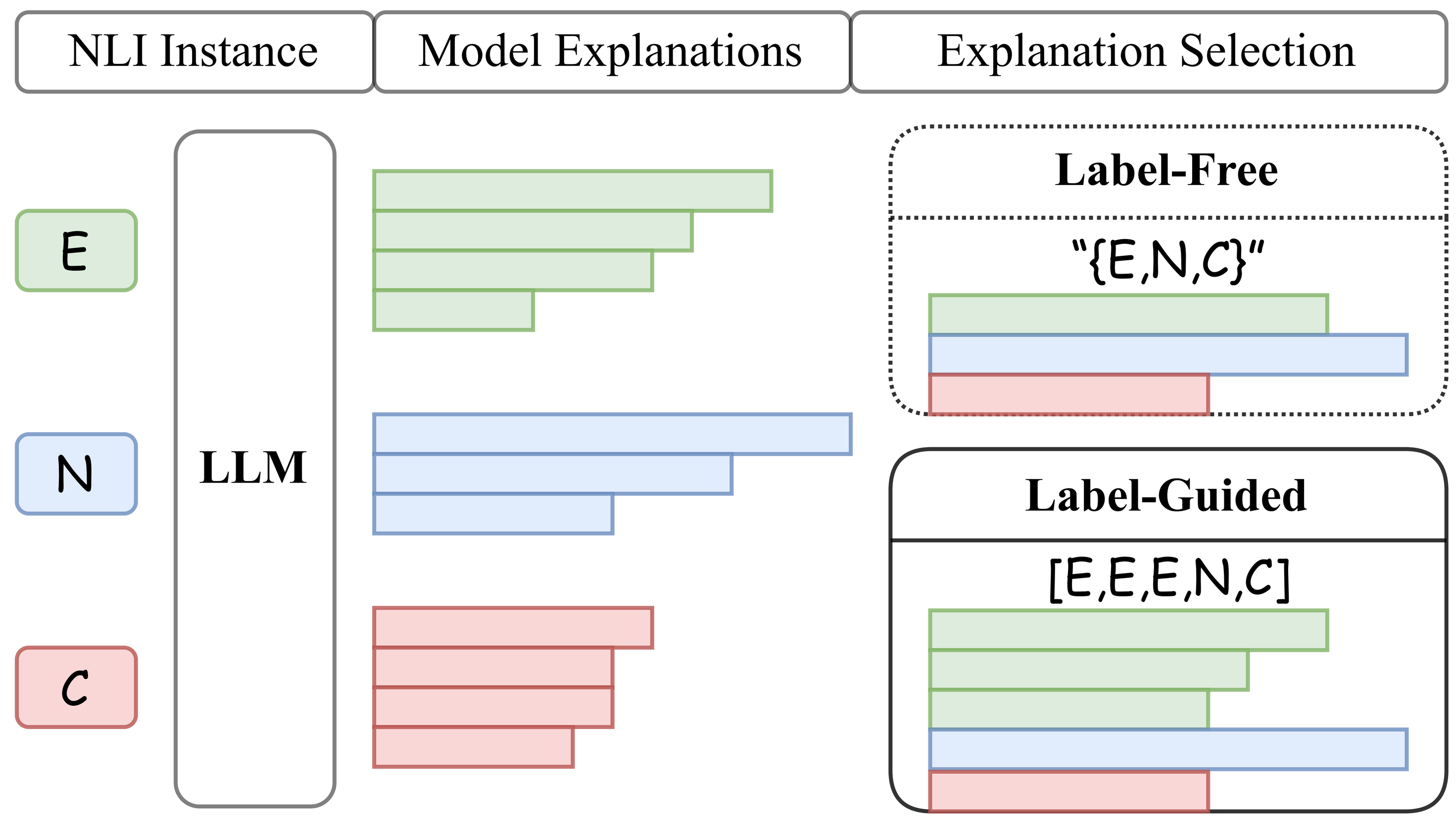
Beiduo Chen [1,2] Siyao Peng [1,2] Anna Korhonen [3] Barbara Plank [1,2]

[1] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] Language Technology Lab, University of Cambridge, United Kingdom

Introduction

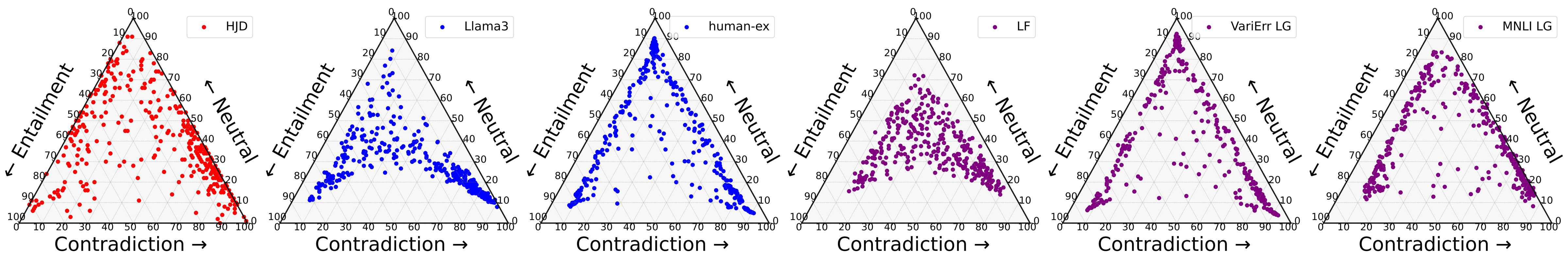


Generating Model Explanations for NLI



Can Model Explanations Help LLMs Approximate HJD as Humans Do?

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison (dev/test)			RoBERTa Fine-Tuning Comparison (dev/test)			Global
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609	0.062 / 0.060	0.933 / 0.922	0.696 / 0.653	1.000
VariErr dist.	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522	0.166 / 0.173	1.246 / 1.261	0.616 / 0.594	0.688
MNLI dist.	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555	0.101 / 0.093	1.052 / 1.020	0.625 / 0.607	0.795
Llama3	0.259	0.262	0.284	0.099 / 0.101	1.045 / 1.044	0.516 / 0.487	0.094 / 0.096	1.030 / 1.031	0.545 / 0.522	0.689
+ human-ex	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.039	0.575 / 0.556	0.091 / 0.092	1.021 / 1.019	0.641 / 0.616	0.771
+ LF model-ex	0.295	0.278	0.310	0.106 / 0.107	1.066 / 1.063	0.539 / 0.533	0.103 / 0.105	1.059 / 1.058	0.581 / 0.571	0.744
+ VariErr LG model-ex	**0.234**	**0.247**	**0.266**	0.097 / 0.098	1.041 / 1.037	0.558 / 0.544	**0.089** / **0.091**	**1.016** / **1.014**	0.633 / 0.626	0.760
+ MNLI LG model-ex	0.242	0.251	0.275	**0.096** / **0.097**	**1.037** / **1.034**	**0.589** / **0.580**	0.090 / 0.092	1.019 / 1.018	**0.657** / **0.645**	**0.849**
GPT-4o	0.265	0.263	0.289	0.103 / 0.096	1.059 / 1.029	0.526 / 0.517	0.093 / 0.092	1.027 / 1.018	0.525 / 0.521	0.703
+ human-ex	**0.187**	**0.207**	**0.223**	0.093 / 0.098	1.027 / 1.036	**0.570** / **0.552**	**0.079** / **0.080**	**0.986** / **0.987**	0.617 / **0.617**	**0.769**
+ LF model-ex	0.252	0.242	0.275	0.101 / 0.102	1.052 / 1.047	0.537 / 0.545	0.157 / 0.167	1.220 / 1.244	0.587 / 0.561	0.752
+ VariErr LG model-ex	0.192	0.209	0.226	**0.092** / **0.093**	**1.026** / **1.022**	0.554 / 0.551	0.088 / 0.089	1.013 / 1.008	**0.618** / 0.598	0.761



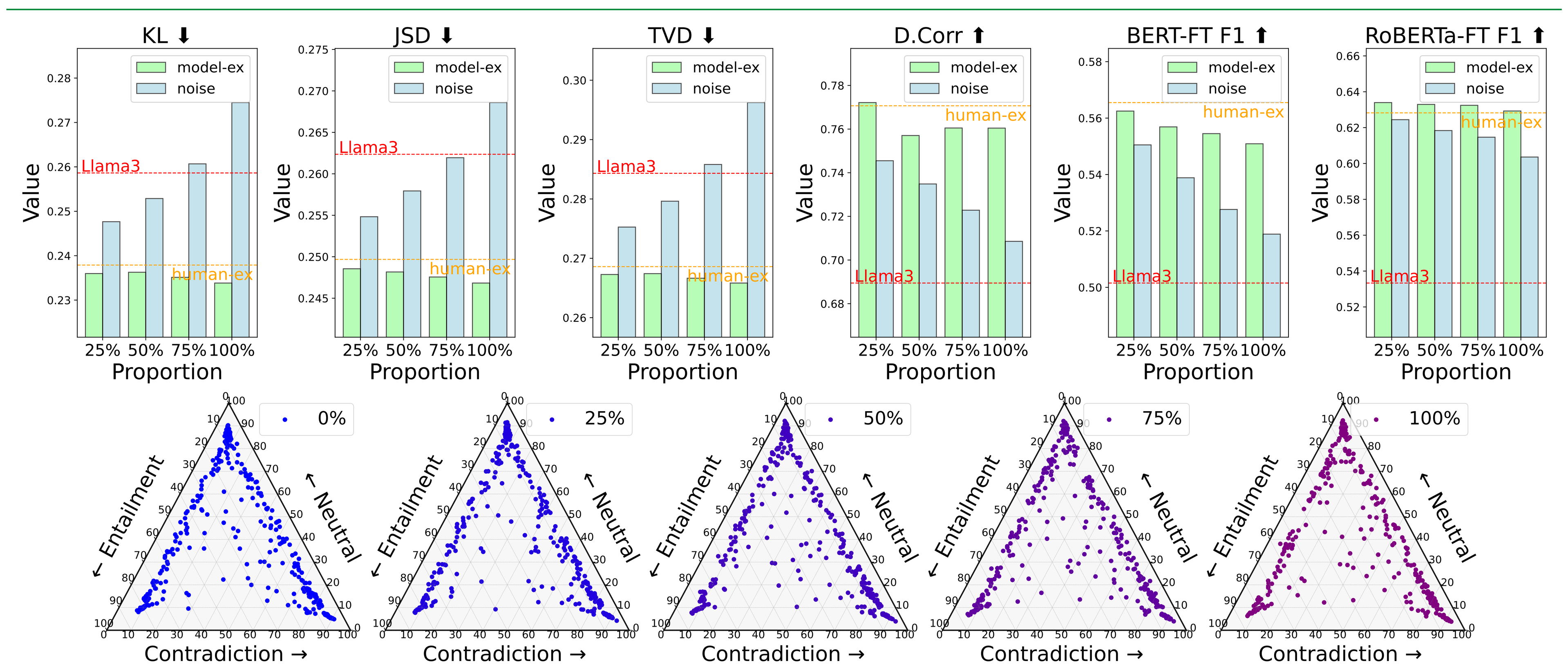
(a) (b) (c) (d) (e) (f)

Can Model-EX Enhance on OOD?

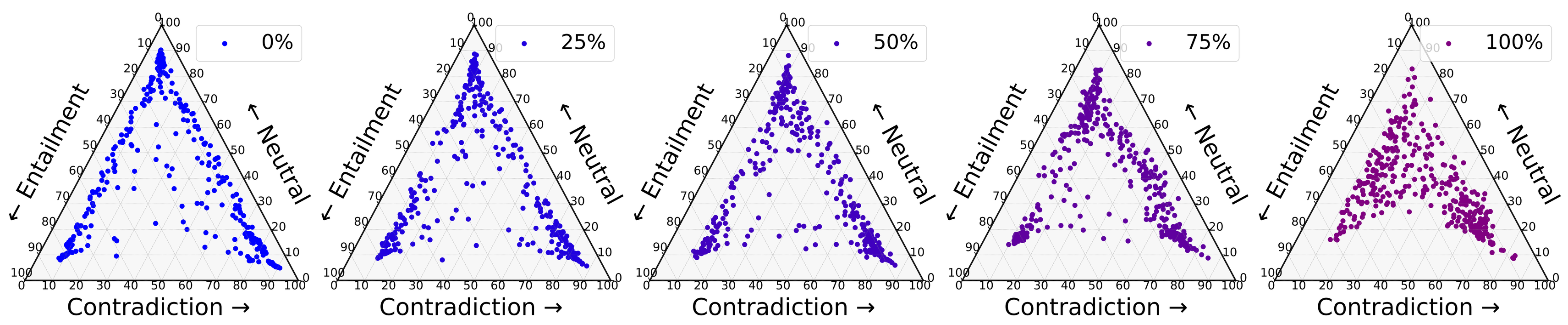
Trained Classifiers	BERT FT Test			RoBERTa FT Test		
	R1 ↑	R2 ↑	R3 ↑	R1 ↑	R2 ↑	R3 ↑
Zero-shot-LM	0.170	0.176	0.197	0.167	0.167	0.168
MNLI-FT-LM	0.220	0.269	0.293	0.292	0.262	0.257
ChaosNLI HJD	0.268	0.289	0.332	0.357	0.331	0.338
VariErr dist	0.302	0.259	0.319	0.402	0.311	0.321
MNLI dist	0.229	0.260	0.279	0.317	0.275	0.281
Llama3	0.246	0.276	0.306	0.304	0.297	0.304
+ human-ex	0.296	0.289	0.349	0.400	**0.330**	**0.344**
+ LF model-ex	0.292	**0.295**	0.328	0.314	0.262	0.323
+ VariErr LG model-ex	**0.305**	0.285	**0.349**	**0.411**	0.324	0.319
+ MNLI LG model-ex	0.284	0.283	0.321	0.339	0.287	0.307
GPT-4o	0.258	0.263	0.295	0.309	0.282	0.302
+ human-ex	**0.351**	**0.294**	**0.332**	**0.393**	**0.324**	**0.325**
+ LF model-ex	0.285	0.283	0.315	0.350	0.282	0.310
+ VariErr LG model-ex	0.341	0.293	0.330	0.393	0.324	0.323

- Model explanations are comparable to humans in approximating HJD on NLI, and can be scaled up from a few annotations of datasets without explanations.
- Modeling HLV information can improve NLI classifiers' performance, and MJDs generated by our method are robust on OOD datasets w/o labels or explanations.

Human versus Model: Are They Different and Does It Matter?



(a) Gradually replaced by **model** explanations.



(b) Gradually replaced by **noise** explanations.

Can Human Preference Lead to Better Selection?

Distributions	Dist. Comparison			RoBERTa Fine-Tuning Comparison(dev/test)			Global
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Llama3	0.258	0.261	0.286	0.092 / 0.095	1.025 / 1.026	0.531 / 0.512	0.684
+ human ex	0.240	0.249	0.275	0.089 / 0.091	1.014 / 1.015	0.618 / 0.597	0.750
+ replace *preferred* model ex							
greedy 75.75%	0.241	0.248	0.274	0.088 / 0.090	1.013 / 1.013	0.619 / 0.594	0.733
representative 55.25%	0.240	0.248	0.274	0.088 / 0.091	1.013 / 1.014	0.619 / 0.597	0.739
+ replace *unpreferred* model ex							
greedy 68.5%	0.239	0.247	0.273	**0.087** / 0.090	**1.011** / 1.012	**0.623** / 0.599	0.752
representative 63.25%	**0.237**	**0.246**	**0.271**	0.088 / **0.090**	1.011 / **1.012**	0.621 / **0.607**	**0.761**

Datasets	Lexical			Syntactic			Semantic		AVG
	n = 1↓	n = 2 ↓	n = 3↓	n = 1↓	n = 2↓	n = 3↓	Cos.↓	Euc.↓	AVG ↓
human-ex	0.335	0.098	0.042	0.767	0.341	0.140	0.528	0.520	0.428
replaced *preferred* model ex									
greedy	0.416	0.157	0.082	0.874	0.488	0.233	0.540	0.532	0.474
represent.	0.392	0.149	0.089	0.835	0.426	0.205	0.542	0.541	0.466
replaced *unpreferred* model ex									
greedy	0.387	0.130	0.069	0.841	0.432	0.196	0.527	0.528	0.457
represent.	0.378	0.130	0.073	0.837	0.426	0.195	0.534	0.532	**0.455**

- Model and human explanations result in similar performance, while noise replacement clearly hurts, indicating that the relevant contents of explanations are crucial
- The potential of *variability* as a metric for measuring the model explanations.

Conclusion

- Experiments show that MJDs from *LLMs and model explanations* result in comparable scores with MJDs from *LLM and human explanations* — A rose by any other name would smell as sweet. (A quote from Romeo and Juliet used to metaphorically argue the intrinsic qualities or nature of something remain the same, regardless of its name or origin.)
- Notably, our approach generalizes to explanation-free datasets and remains effective in challenging out-of-domain test sets. Results indicate that LLM-generated explanations can significantly reduce annotation costs, making it a scalable and efficient proxy for capturing human label variation.

Resource

Paper Code