Understanding and Modeling Human Label Variation in LLM

—*Natural Language Inference as A Case*

Beiduo Chen

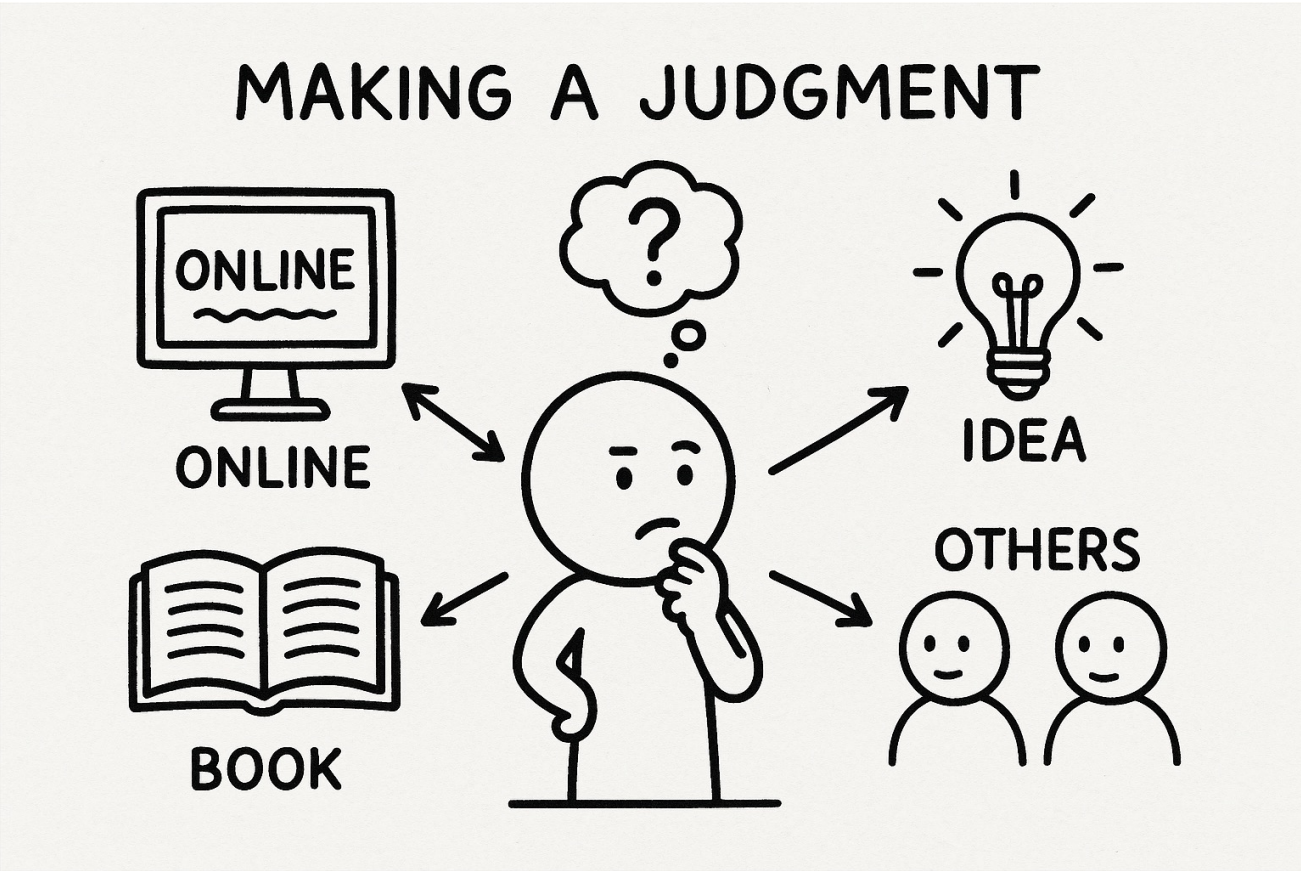
MaiNLP lab, LMU Munich

LTL lab, University of Cambridge

WDMD 2025: 4th International Workshop on Dependability Modeling and Digitalization

June 25, 2025

How Do You Make a Judment?



What is Human Label Variation?

The "Problem" of Human Label Variation:
On Ground Truth in Data, Modeling and Evaluation

Journal of Artificial Intelligence Research 72 (2021) 1385-1470 Submitted 02/2021; published 12/2021

Center for Information and Lang
Munich Center for

Data

Hu

Learning from Disagreement: A Survey

Alexandra N. Uma A.N.UMA@QMUL.AC.UK
Queen Mary University of London

Tommaso Fornaciari FORNACIARI.TOMMASO@UNIBOCCONI.IT
Dirk Hovy DIRK.HOVY@UNIBOCCONI.IT
Università B

Silviu Pau
Queen Mary

Barbara P
IT Universit

Massimo F
Queen Mary

Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning

Tommaso Fornaciari
Università Bocconi
fornaciari@unibocconi.it

Alexandra Uma
Queen Mary University
a.n.uma@qmul.ac.uk

Silviu Paun
Queen Mary University
s.paun@qmul.ac.uk

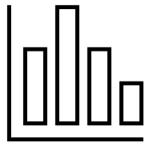
Barbara Plank
IT University of Copenhagen
bapl@itu.dk

Dirk Hovy
Università Bocconi
dirk.hovy@unibocconi.it

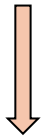
Massimo Poesio
Queen Mary University
m.poesio@qmul.ac.uk

To Utilize Human Label Variation (HLV)

{“A”: 2,
“B”: 3,
“C”: 2,
“D”: 1}



soft labels



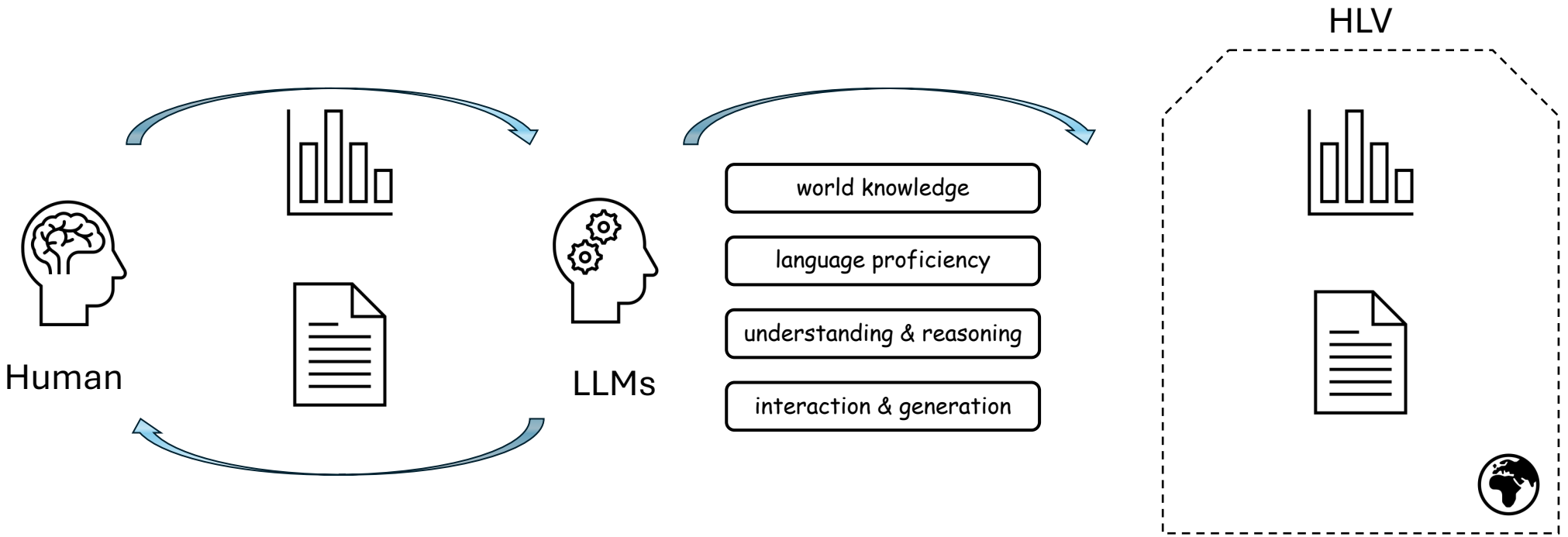
Language Model

Implicitly convey HLV information

- cannot capture the reasons behind

- difficult to further analyze HLV

Large Language Models (LLMs)



HLV Matters for LLMs

- Real-world data is ambiguous and multi-faceted
- HLV provides richer signal for training and evaluation
- Ignoring it leads to:
  - Overconfidence
  - Reduced robustness
- Embracing HLV = Better uncertainty modeling & reliability

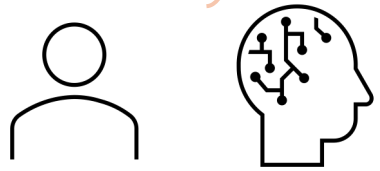
Natural Language Inference as A Case

Explainable

inference and reasoning tasks, where the questions have a relatively objective nature

comprehensive and unbiased understanding

Explainable **easier to interpret** Explanations



RQ1: Can LLM achieve approximating human label distributions when provided with explanations?

RQ2: How to design the evaluation for assessing the effectiveness of this approximation?

RQ3: How can we employ LLMs to improve their simulation ability without human explanations?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

"Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations?

Findings of the Association for Computational Linguistics: EMNLP 2024

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, Barbara Plank

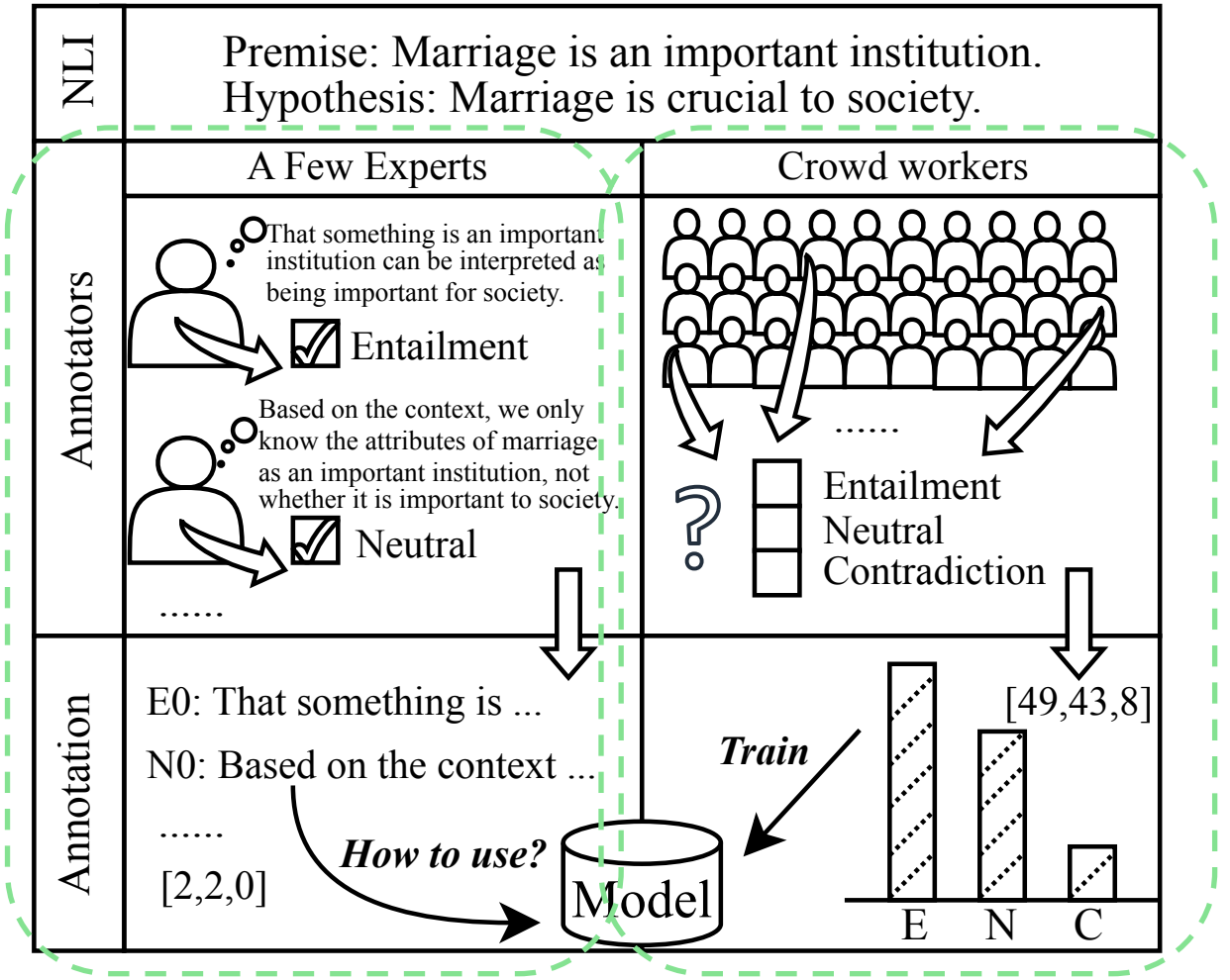
RQ1: Can LLM achieve approximating human label distributions when provided with explanations?

RQ2: How to design the evaluation for assessing the effectiveness of this approximation?






Introduction



Research Questions

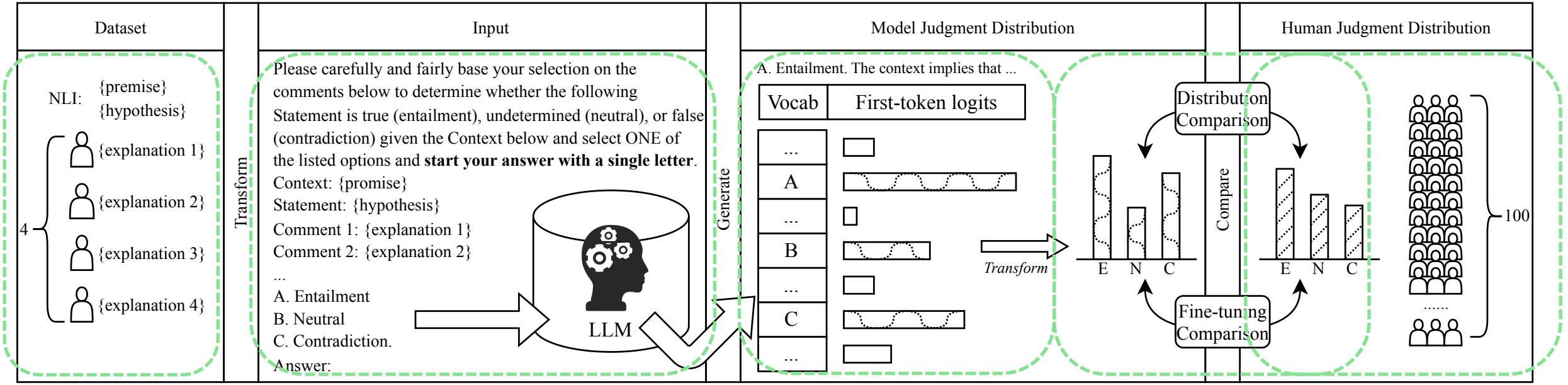
- Can LLMs provided with a "small" number of detailed explanations better approximate the human judgment distributions collected by a "big" number of annotators?

RQ1: Can LLM achieve approximating human label distributions when provided with explanations?

- Are the obtained model judgment distributions (MJDs) suitable as soft labels for fine-tuning smaller models to predict distributions?

RQ2: How to design the evaluation for assessing the effectiveness of this approximation?

Method


Dataset
Input
Model Judgment Distribution
Human Judgment Distribution
NLI:
{premise}
{hypothesis}
4
{explanation 1}
{explanation 2}
{explanation 3}
{explanation 4}
Transform
Please carefully and fairly base your selection on the comments below to determine whether the following Statement is true (entailment), undetermined (neutral), or false (contradiction) given the Context below and select ONE of the listed options and **start your answer with a single letter**.
Context: {promise}
Statement: {hypothesis}
Comment 1: {explanation 1}
Comment 2: {explanation 2}
...
A. Entailment
B. Neutral
C. Contradiction.
Answer:
LLM
Generate
A. Entailment. The context implies that ...
Vocab
First-token logits
...
A
...
B
...
C
...
Transform
Distribution Comparison
Compare
Fine-tuning Comparison
E N C
E N C
100
.....

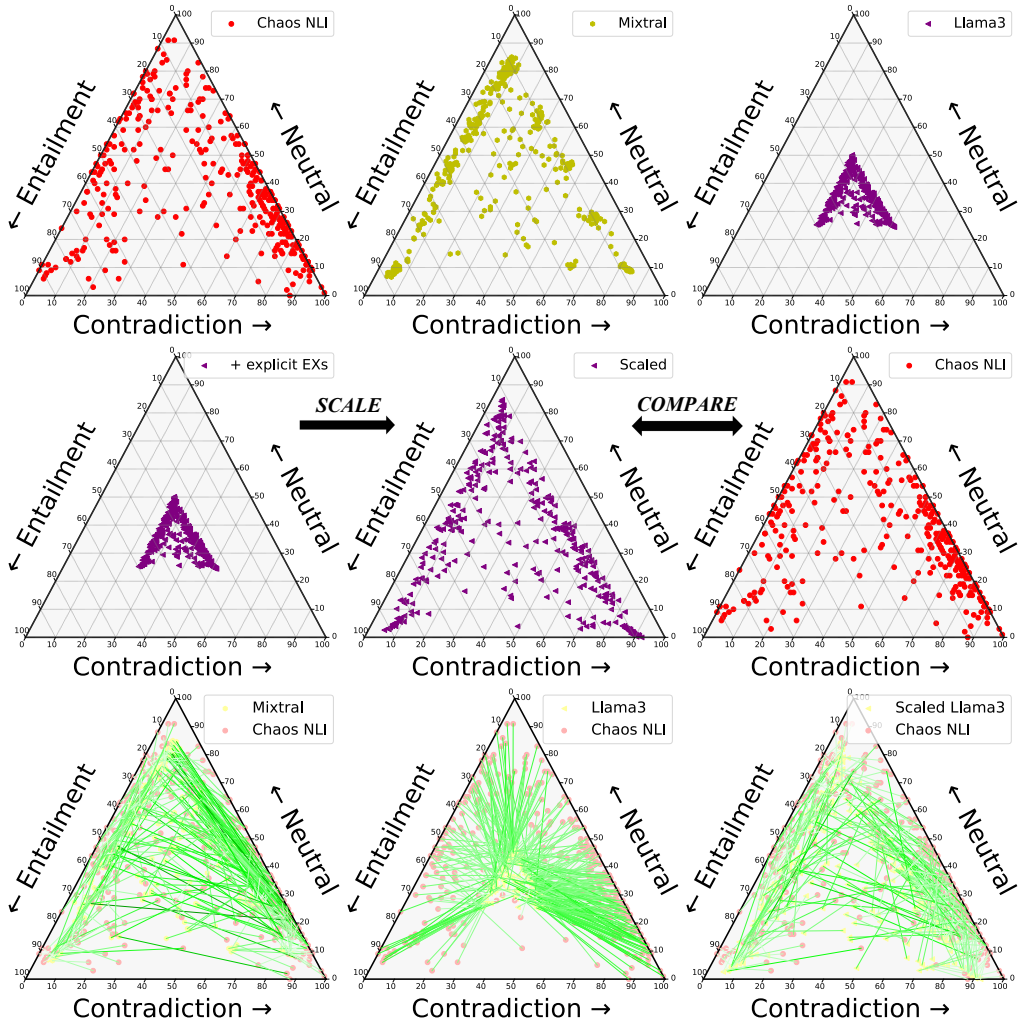
Results

Distributions\Metrics	KL ↓	JSD ↓	TVD ↓
Baseline			
Chaos NLI	0	0	0
MNLI single label	9.288	0.422	0.435
MNLI distributions	1.242	0.281	0.295
VariErr distributions	3.604	0.282	0.296
Uniform distribution	0.364	0.307	0.350
MJDs from Mixtral			
p_{norm} of Mixtral	0.433	0.291	0.340
+ "serial" explanations	0.407	0.265	0.306
+ "serial" explicit explanations	0.382	0.246	0.286
+ "parallel" explanations	0.339	0.258	0.295
+ "parallel" explicit explanations	**0.245**	**0.211**	**0.239**
p_{sfmax} of Mixtral	0.434	0.292	0.342
+ "serial" explanations	0.349	0.258	0.296
+ "serial" explicit explanations	0.305	0.235	0.269
+ "parallel" explanations	0.310	0.255	0.290
+ "parallel" explicit explanations	**0.217**	**0.208**	**0.232**
MJDs from Llama3			
p_{norm} of Llama3	0.259	0.262	0.284
+ "serial" explanations	0.255	0.259	0.281
+ "serial" explicit explanations	**0.235**	**0.247**	**0.266**
+ "parallel" explanations	0.257	0.261	0.283
+ "parallel" explicit explanations	0.243	0.253	0.273
p_{sfmax} of Llama3	0.231	0.245	0.260
+ "serial" explanations	0.226	0.243	0.258
+ "serial" explicit explanations	**0.212**	**0.232**	**0.245**
+ "parallel" explanations	0.226	0.245	0.260
+ "parallel" explicit explanations	0.214	0.237	0.254

inconsistent

Distributions	BERT FT (dev / test)			RoBERTa FT (dev / test)		
	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓
Baseline						
Chaos NLI train set	**0.626 / 0.646**	**0.074 / 0.077**	**0.972 / 0.974**	**0.699 / 0.650**	**0.061 / 0.067**	**0.932 / 0.943**
MNLI single label	0.561 / 0.589	0.665 / 0.704	2.743 / 2.855	0.635 / 0.603	0.844 / 0.867	3.281 / 3.344
MNLI distributions	0.546 / 0.543	0.099 / 0.102	1.046 / 1.048	0.613 / 0.604	0.100 / 0.096	1.047 / 1.029
VariErr distributions	0.557 / 0.559	0.179 / 0.186	1.286 / 1.299	0.617 / 0.589	0.174 / 0.197	1.269 / 1.333
MJDs from Mixtral						
p_{norm} of Mixtral	0.416 / 0.422	0.134 / 0.133	1.152 / 1.142	0.486 / 0.466	0.123 / 0.127	1.118 / 1.123
+ "serial" explanations	0.443 / 0.454	0.145 / 0.141	1.183 / 1.166	0.509 / 0.514	0.128 / 0.128	1.132 / 1.126
+ "serial" explicit explanations	0.506 / 0.511	0.130 /0.130	1.139 / 1.132	**0.569 / 0.572**	0.114 / 0.122	1.091 / 1.107
+ "parallel" explanations	0.404 / 0.428	0.134 / 0.131	1.150 / 1.136	0.483 / 0.502	0.123 / 0.122	1.118 / 1.109
+ "parallel" explicit explanations	**0.507 / 0.514**	**0.108 / 0.108**	**1.074 / 1.065**	0.558 / 0.565	**0.092 / 0.098**	**1.025 / 1.037**
p_{sfmax} of Mixtral	0.427 / 0.432	0.131 / 0.129	1.140 / 1.130	0.497 / 0.472	0.121 / 0.125	1.112 / 1.118
+ "serial" explanations	0.452 / 0.462	0.121 / 0.118	1.113 / 1.096	0.506 / 0.525	0.110 / 0.109	1.078 / 1.069
+ "serial" explicit explanations	0.509 / **0.520**	0.105 / 0.105	1.064 / 1.057	**0.568** / 0.573	0.093 / 0.098	1.026 / 1.036
+ "parallel" explanations	0.397 / 0.429	0.121 / 0.119	1.112 / 1.098	0.497 / 0.505	0.110 / 0.111	1.079 / 1.074
+ "parallel" explicit explanations	**0.522** / 0.517	**0.095 / 0.095**	**1.035 / 1.026**	0.567 / **0.576**	**0.082 / 0.087**	**0.994 / 1.003**
MJDs from Llama3						
p_{norm} of Llama3	0.514 / 0.526	0.097 / 0.098	1.038 / 1.036	0.541 / 0.528	0.091 / 0.094	1.023 / 1.025
+ "serial" explanations	0.574 / 0.574	0.096 / 0.097	1.037 / 1.033	0.618 / 0.601	0.091 / 0.093	1.020 / 1.022
+ "serial" explicit explanations	0.578 / 0.574	**0.091 / 0.092**	**1.022 / 1.018**	0.634 / 0.598	**0.085 / 0.088**	**1.003 / 1.006**
+ "parallel" explanations	0.573 / 0.582	0.098 / 0.098	1.041 / 1.038	0.636 / 0.598	0.093 / 0.095	1.026 / 1.028
+ "parallel" explicit explanations	**0.582 / 0.586**	0.094 / 0.095	1.030 / 1.026	**0.639 / 0.620**	0.089 / 0.091	1.014 / 1.016
p_{sfmax} of Llama3	0.528 / 0.524	0.091 / 0.093	1.023 / 1.021	0.546 / 0.535	0.085 / 0.089	1.005 / 1.009
+ "serial" explanations	0.567 / 0.576	0.091 / 0.091	1.021 / 1.016	0.626 / 0.608	0.082 / 0.086	0.996 / 1.000
+ "serial" explicit explanations	**0.585** / 0.568	**0.086 / 0.087**	**1.008 / 1.004**	**0.646** / 0.610	**0.077 / 0.081**	**0.981 / 0.987**
+ "parallel" explanations	0.584 / **0.583**	0.092 / 0.093	1.024 / 1.020	0.643 / 0.611	0.085 / 0.089	1.004 / 1.008
+ "parallel" explicit explanations	0.581 / 0.578	0.088 / 0.089	1.014 / 1.010	0.645 / **0.621**	0.081 / 0.085	0.993 / 0.996

Results



align

Distributions\Metrics	D.Corr ↑
Uniform distribution	0
MNLI single label	0.612
MNLI distributions	0.795
VariErr distributions	0.688
MJDs from Mixtral	
p_{norm} of Mixtral	0.609
+ "parallel" explicit explanations	**0.719**
p_{sfmax} of Mixtral	0.593
+ "parallel" explicit explanations	**0.709**
MJDs from Llama3	
p_{norm} of Llama3	0.689
+ "parallel" explicit explanations	**0.809**
p_{sfmax} of Llama3	0.677
+ "parallel" explicit explanations	**0.802**

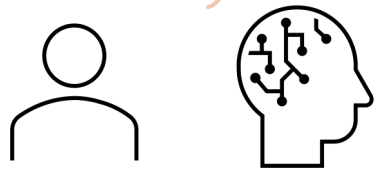
Natural Language Inference as A Case

Explainable

inference and reasoning tasks, where the questions have a relatively objective nature

comprehensive and unbiased understanding

Explainable **easier to interpret** Explanations



RQ1: Can LLM achieve approximating human label distributions when provided with explanations?

RQ2: How to design the evaluation for assessing the effectiveness of this approximation?

RQ3: How can we employ LLMs to improve their simulation ability without human explanations?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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

Findings of the Association for Computational Linguistics: ACL 2025

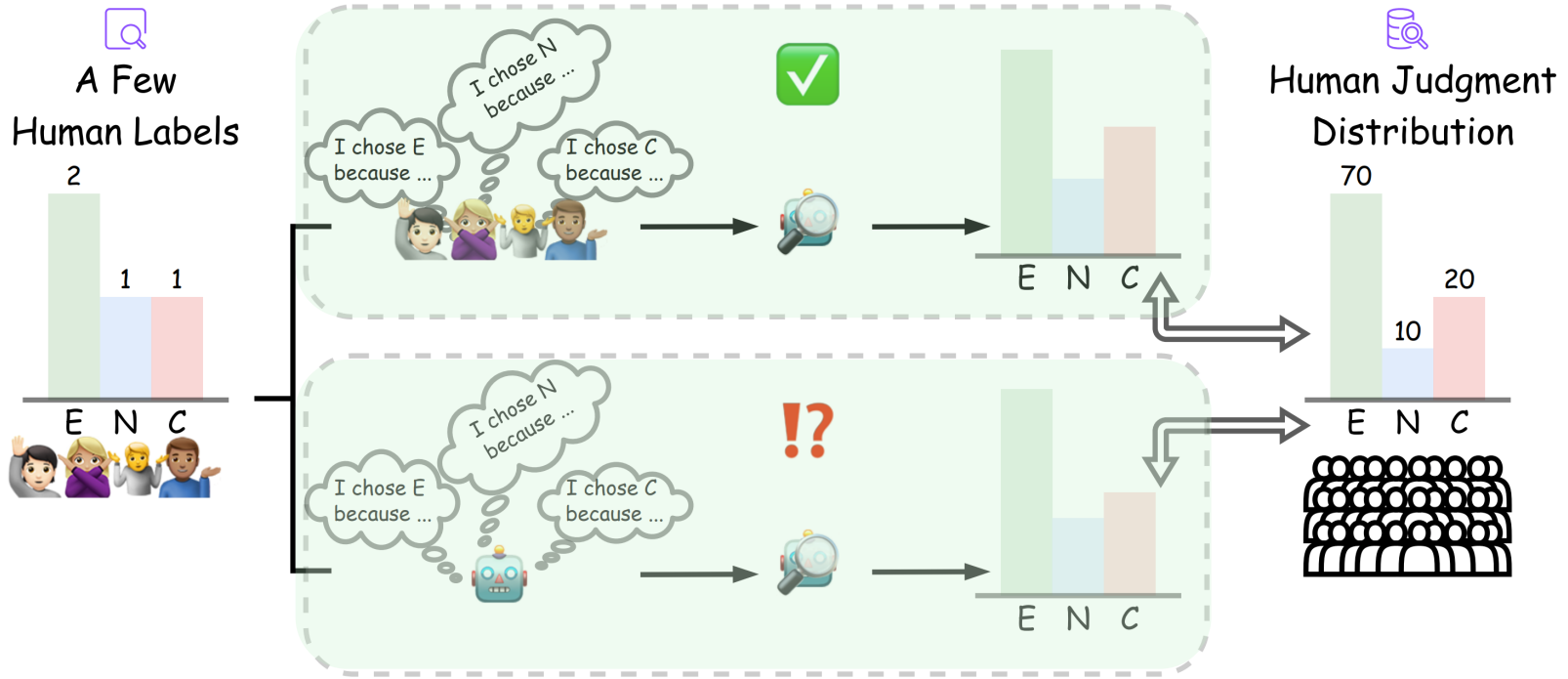
Beiduo Chen, Siyao Peng, Anna Korhonen, Barbara Plank

RQ2: How to design the evaluation for assessing the effectiveness of this approximation?

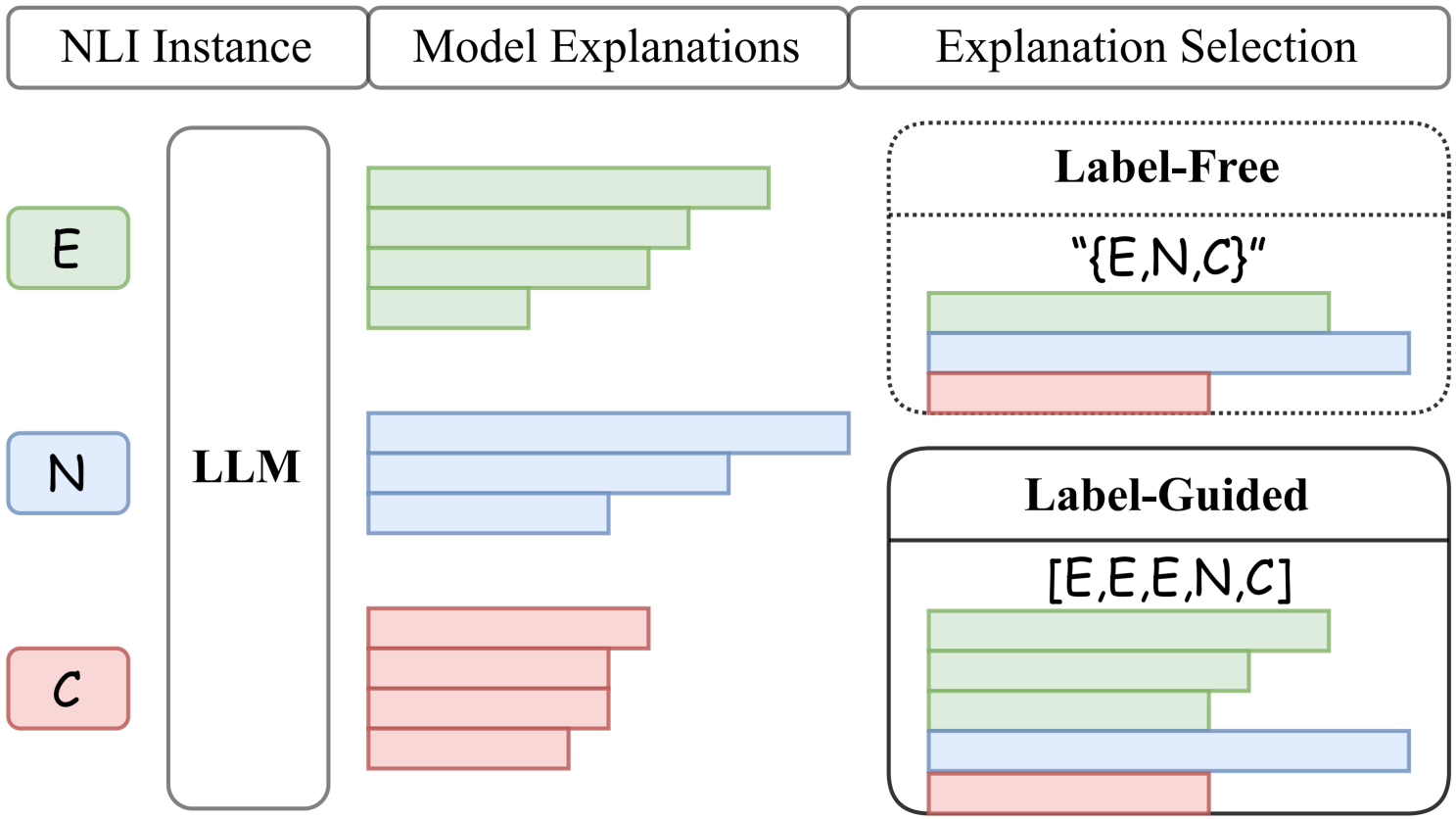
RQ3: How can we employ LLMs to improve their simulation ability without human explanations?




Introduction

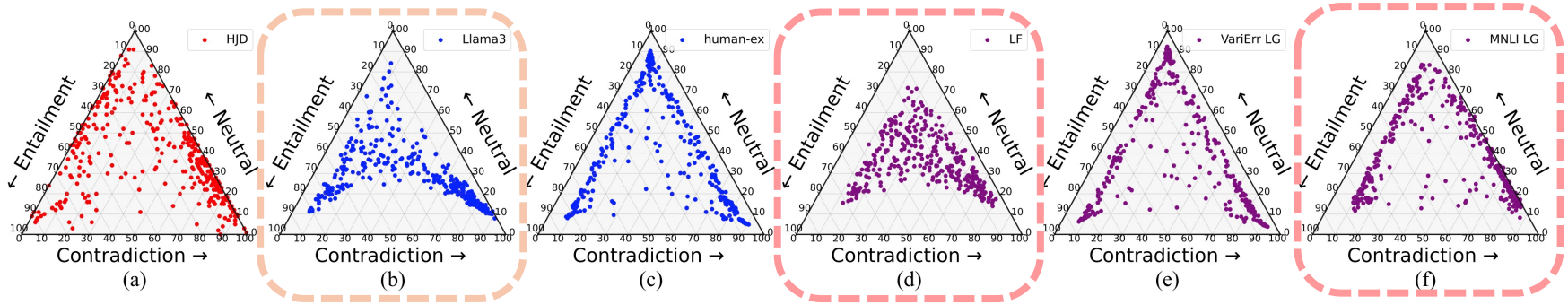


Method


NLI Instance
Model Explanations
Explanation Selection
E
N
C
LLM
Label-Free
"{E,N,C}"
Label-Guided
[E,E,E,N,C]

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 Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
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 distribution	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 distribution	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
Model Judgment Distributions										
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 explanations	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
+ model explanations										
Label-Free	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 Label-Guided	**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 Label-Guided	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 explanations	**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**
+ model explanations										
Label-Free	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 Label-Guided	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



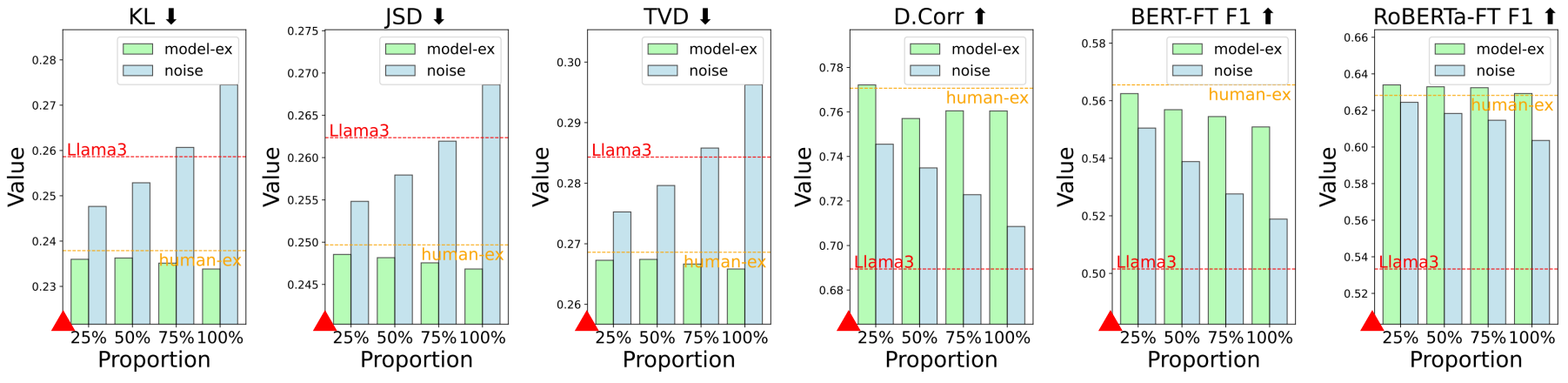
Can Model-Generated Explanations Enhance Performance on OOD Task?

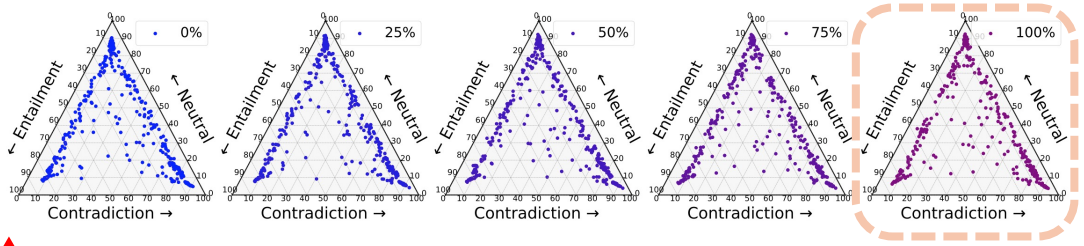
Classifiers	BERT FT Test			RoBERTa FT Test		
	R1 ↑	R2 ↑	R3 ↑	R1 ↑	R2 ↑	R3 ↑
Classifiers without distribution training						
Out-of-the-box LM	0.170	0.176	0.197	0.167	0.167	0.168
MNLI-FT-LM	0.220	0.269	0.293	0.292	0.262	0.257
Classifiers trained on label distributions						
ChaosNLI HJD	0.268	0.289	0.332	0.357	0.331	0.338
VariErr distribution	0.302	0.259	0.319	0.402	0.311	0.321
MNLI distribution	0.229	0.260	0.279	0.317	0.275	0.281
Classifiers trained on MJDs						
Llama3	0.246	0.276	0.306	0.304	0.297	0.304
+ human explanations	0.296	0.289	0.349	0.400	**0.330**	**0.344**
+ model explanations						
Label-Free	0.292	**0.295**	0.328	0.314	0.262	0.323
VariErr Label-Guided	**0.305**	0.285	**0.349**	**0.411**	0.324	0.319
MNLI Label-Guided	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 explanations	**0.351**	**0.294**	**0.332**	**0.393**	**0.324**	**0.325**
+ model explanations						
Label-Free	0.285	0.283	0.315	0.350	0.282	0.310
VariErr Label-Guided	0.341	0.293	0.330	0.393	0.324	0.323

HLV works

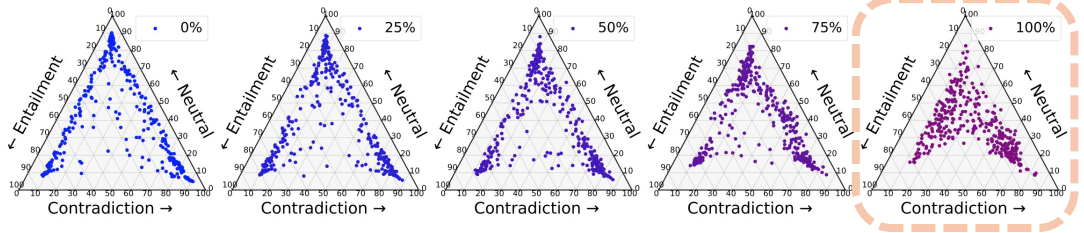
MJD works

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





(a) Gradually replaced by model explanations.



(b) Gradually replaced by noise explanations.

Distributions	Dist. Comparison			Global Metric
	KL ↓	JSD ↓	TVD ↓	D.Corr ↑
VariErr distribution	6.628	0.357	0.352	0.907
Llama3 MJD	0.029	0.068	0.088	0.691
+ human explanations	0.000	0.000	0.000	1.000
+ replace model explanations				
Label-Free 100%	0.024	0.067	0.088	0.647
VariErr Label-Guided 25%	**0.001**	**0.012**	**0.015**	**0.977**
VariErr Label-Guided 50%	0.003	0.017	0.022	0.959
VariErr Label-Guided 75%	0.003	0.019	0.024	0.950
VariErr Label-Guided 100%	0.004	0.021	0.027	0.939

Can Human Preference Lead to Better Explanation Selection?

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison(dev/test)			RoBERTa Fine-Tuning Comparison(dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Llama3	0.258	0.261	0.286	0.092 / 0.093	1.024 / 1.020	0.514 / 0.471	0.092 / 0.095	1.025 / 1.026	0.531 / 0.512	0.684
+ human explanations	0.240	0.249	0.275	0.090 / 0.090	1.017 / 1.011	0.594 / 0.567	0.089 / 0.091	1.014 / 1.015	0.618 / 0.597	0.750
+ replace *preferred* model explanations										
greedy 75.75%	0.241	0.248	0.274	0.089 / 0.090	1.017 / 1.011	0.584 / 0.569	0.088 / 0.090	1.013 / 1.013	0.619 / 0.594	0.733
representative 55.25%	0.240	0.248	0.274	0.089 / 0.090	1.016 / 1.011	0.587 / 0.567	0.088 / 0.091	1.013 / 1.014	0.619 / 0.597	0.739
+ replace *unpreferred* model explanations										
greedy 68.5%	0.239	0.247	0.273	**0.089 / 0.089**	**1.016 / 1.009**	**0.589 / 0.571**	**0.087** / 0.090	**1.011** / 1.012	**0.623** / 0.599	0.752
representative 63.25%	**0.237**	**0.246**	**0.271**	0.089 / 0.089	1.016 / 1.010	0.584 / 0.566	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 explanations									
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 explanations									
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**

Variance Within a Set of Explanations

Within Label Variation

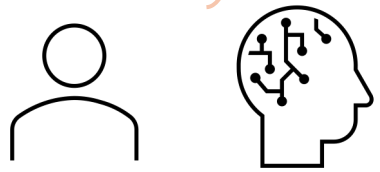
Natural Language Inference as A Case

Explainable

inference and reasoning tasks, where the questions have a relatively objective nature

comprehensive and unbiased understanding

Explainable **easier to interpret** Explanations



- RQ1: Can LLM achieve approximating human label distributions when provided with explanations?
- RQ2: How to design the evaluation for assessing the effectiveness of this approximation?
- RQ3: How can we employ LLMs to improve their simulation ability without human explanations?

Conclusions & Takeaways

- Human disagreement is inevitable—and valuable
  - Train more robust models
  - Better reflect human uncertainty
- LLMs + explanations (human or model) = scalable HLV modeling
  - EMNLP24: Use few human explanations to approximate HJD
  - ACL25: Replace human explanations with LLM-generated ones
- Reduces cost of soft label supervision
- Supports more **reliable**, **transparent**, and **human-aligned** AI systems

Thank You !

Beiduo Chen
beiduochen@cis.lmu.de
MaiNLP lab, LMU Munich
LTL lab, University of Cambridge

WDMD 2025: 4th International Workshop on Dependability Modeling and Digitalization
June 25, 2025