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

Beiduo Chen [1,2] Xinpeng Wang [1,2] Siyao Peng [1,2] Robert Litschko [1,2] Anna Korhonen [3] Barbara Plank [1,2]

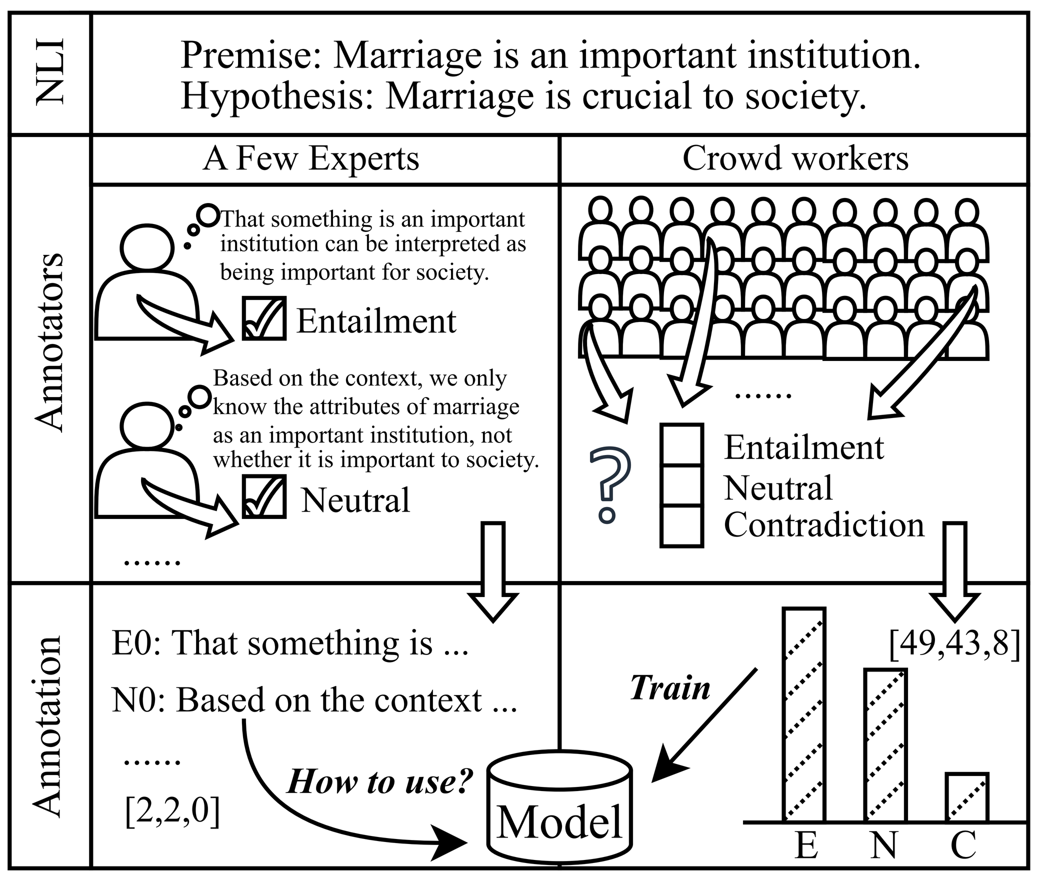
[1] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] Language Technology Lab, University of Cambridge, United Kingdom

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN · UNIVERSITY OF CAMBRIDGE · Munich Center for Machine Learning

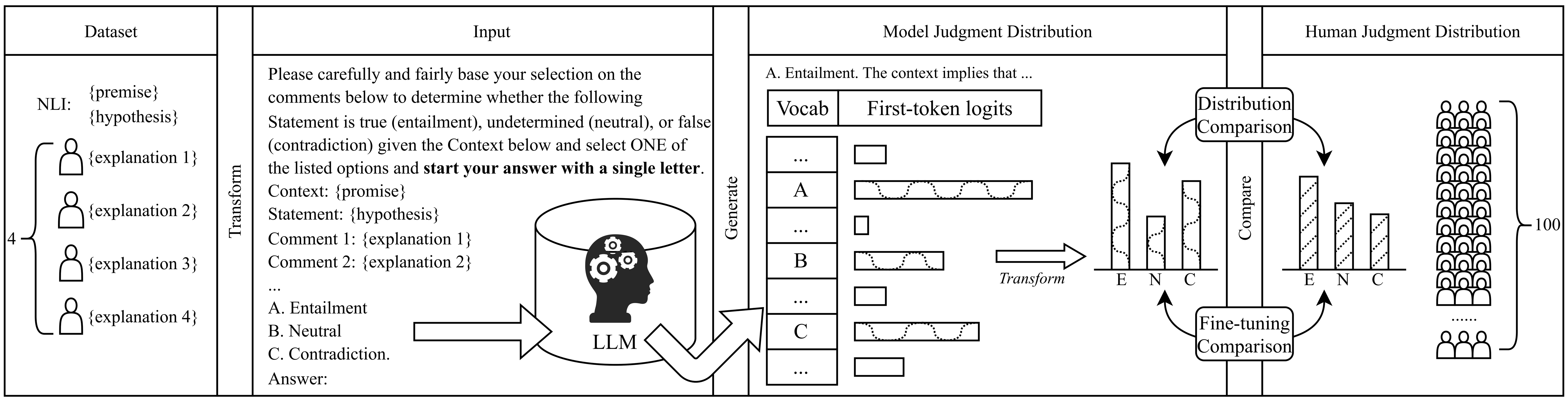
Introduction

- Object: **Human Label Variation** (HLV) is a valuable source of information that arises when multiple human annotators provide different labels for valid reasons.
- Task: In **Natural Language Inference**, approaches to capturing HLV involve either collecting annotations from many _crowd workers_ to represent _human judgment distribution_ (HJD) or use _expert linguists_ to provide detailed _explanations_ for their chosen labels.
- Tool: **Large Language Models** (LLMs) are increasingly used as evaluators (“LLM judges”) but with mixed results, and few works aim to study HJDs.
- Question: *1. Can LLMs provided with a “small” number of detailed explanations better approximate the human judgment distributions collected by a “big” number of annotators?*
  2. Are the obtained model judgment distributions (MJDs) suitable as soft labels for fine-tuning smaller models to predict distributions?

Investigate HLV in NLI



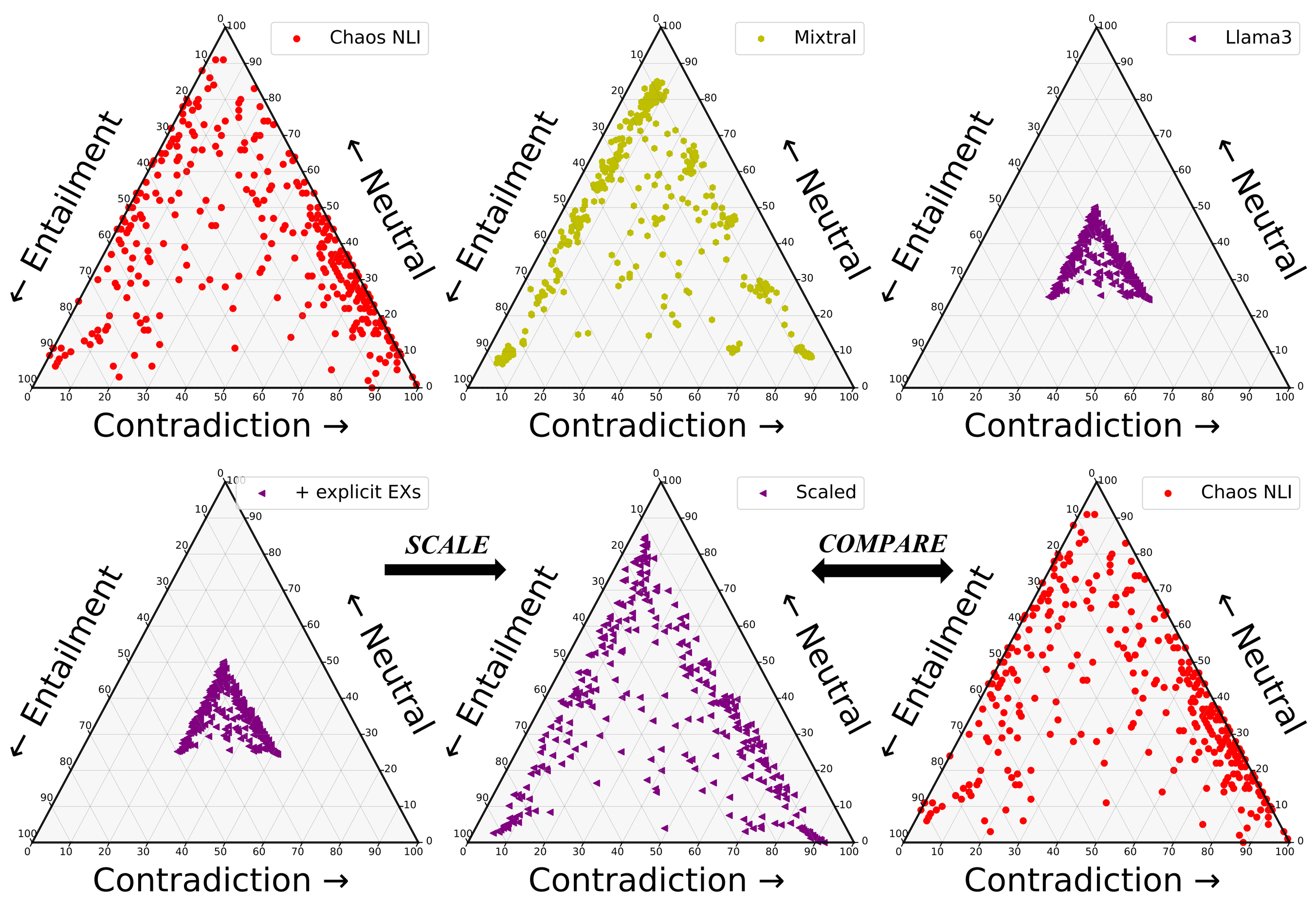
LLMs to Estimate Human Judgment Distributions



Experiments

Distributions	Dist. Comparison KL ↓	Dist. Comparison JSD ↓	Dist. Comparison TVD ↓	BERT Fine-Tuning Comparison(dev/test) Weighted F1 ↑	BERT KL ↓	BERT CE Loss ↓	RoBERTa Fine-Tuning Comparison(dev/test) Weighted F1 ↑	RoBERTa KL ↓	RoBERTa CE Loss ↓	Global Metric D.Corr ↑
Chaos NLI	0	0	0	0.626 / 0.646	0.074 / 0.077	0.972 / 0.974	0.699 / 0.650	0.061 / 0.067	0.932 / 0.943	1
MNLI one-hot	9.288	0.422	0.435	0.561 / 0.589	0.665 / 0.704	2.743 / 2.855	0.635 / 0.603	0.844 / 0.867	3.281 / 3.344	0.612
MNLI dist.	1.242	0.281	0.295	0.546 / 0.543	0.099 / 0.102	1.046 / 1.048	0.613 / 0.604	0.100 / 0.096	1.047 / 1.029	0.795
VariErr dist.	3.604	0.282	0.296	0.557 / 0.559	0.179 / 0.186	1.286 / 1.299	0.617 / 0.589	0.174 / 0.197	1.269 / 1.333	0.688
Uniform dist.	0.364	0.307	0.350	-	-	-	-	-	-	0
$\boldsymbol{p}_{\text{norm}}$ of Mixtral	0.433	0.291	0.340	0.416 / 0.422	0.134 / 0.133	1.152 / 1.142	0.486 / 0.466	0.123 / 0.127	1.118 / 1.123	0.609
+ explanations	0.245	0.211	0.239	0.507 / 0.514	0.108 / 0.108	1.074 / 1.065	**0.569** / 0.572	0.092 / 0.098	1.025 / 1.037	**0.719**
$\boldsymbol{p}_{\text{sfmax}}$ of Mixtral	0.434	0.292	0.342	0.427 / 0.432	0.131 / 0.129	1.140 / 1.130	0.497 / 0.472	0.121 / 0.125	1.112 / 1.118	0.593
+ explanations	**0.217**	**0.208**	**0.232**	**0.522** / **0.520**	**0.095** / **0.095**	**1.035** / **1.026**	0.568 / **0.576**	**0.082** / **0.087**	**0.994** / **1.003**	0.709
$\boldsymbol{p}_{\text{norm}}$ of Llama3	0.259	0.262	0.284	0.514 / 0.526	0.097 / 0.098	1.038 / 1.036	0.541 / 0.528	0.091 / 0.094	1.023 / 1.025	0.689
+ explanations	0.235	0.247	0.266	0.582 / **0.586**	0.091 / 0.092	1.022 / 1.018	0.639 / 0.620	0.085 / 0.088	1.003 / 1.006	**0.809**
$\boldsymbol{p}_{\text{sfmax}}$ of Llama3	0.231	0.245	0.260	0.528 / 0.524	0.091 / 0.093	1.023 / 1.021	0.546 / 0.535	0.085 / 0.089	1.005 / 1.009	0.677
+ explanations	**0.212**	**0.232**	**0.245**	**0.585** /0.583	**0.086** / **0.087**	**1.008** / **1.004**	**0.646** / **0.621**	**0.077** / **0.081**	**0.981** / **0.987**	0.802

Ternary Visualization



Discussion

- FT Comparison _cannot_ be predicted well by Dist. Comparison.
- Llama3 and Mixtral exhibit rather **different clusters**. However, further **zooming in** on Llama3 MJD shows that Llama3 is slightly skewed towards the right side (Contradiction), more _in line with_ Chaos NLI, which _corroborates_ Llama’s superior performance in FT Comparison.
- **Distance Correlation** proves Llama3 is _globally_ better.
- Instance-level metrics are better _complemented_ by additional investigations on the **shape** and **smoothness** of the resulting annotations using _visualization_ and _global_ measures.
- We encourage an uptake of **explanation-informed** datasets.

Resource

Paper

Code

EMNLP 2024 *The 2024 Conference on Empirical Methods in Natural Language Processing*

ellis · erc European Research Council Established by the European Commission · UK Research and Innovation