On Human Label Variation and Explanations in Natural Language Inference

Dr. Siyao (Logan) Peng

MaiNLP, LMU Munich & MCML

25.09.2025 @ MCML-LAMARR workshop, Bonn, Germany

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

fnrs LA LIBERTÉ DE CHERCHER

CIS

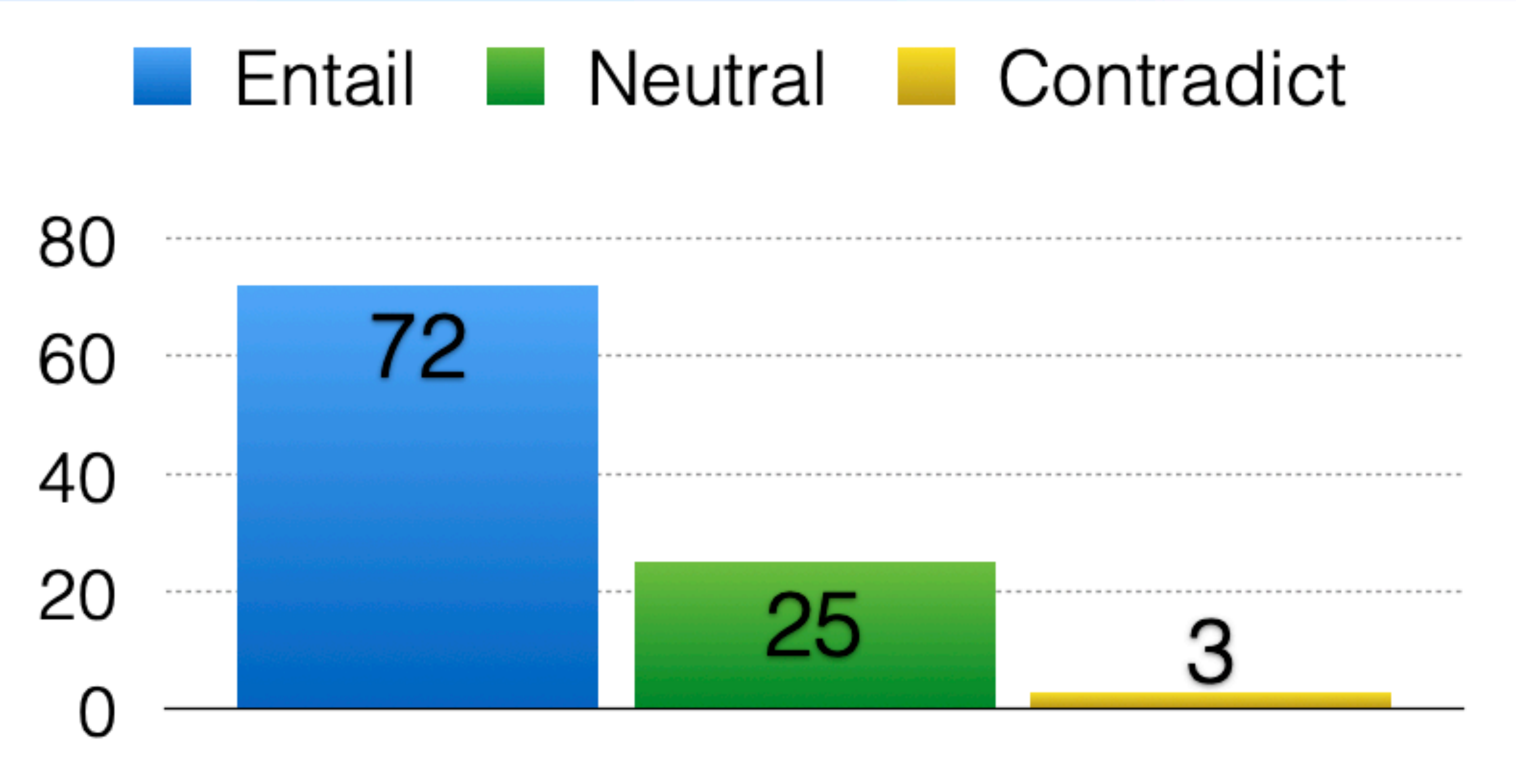
Munich Center for Machine Learning

On many NLP tasks, **more than a single label is plausible.**

Premise: As he stepped across the threshold,
Tommy brought the picture down with terrific force on his head.

Hypothesis: Tommy hurt his head bringing the picture down.

source: 77893n from MNLI (Williams et al., 2018)



Distribution from ChaosNLI
(Nie et al. 2020)

Despite label variation, annotators may vary in their text understanding **when giving the same label.**

Premise: A man in an Alaska sweatshirt stands behind a counter.
Hypothesis: The man is wearing a tank top.
Label: Contradiction
Source: SNLI

Explanation 1: The man cannot simultaneously be wearing a sweatshirt and a tank top.

Explanation 2: A man in Alaska would typically not be wearing a tank top, as it is rather cold there most times of the year.

In this talk, we will briefly discuss:

1. **How to Separate Annotation Error from Human Label Variation?** (ACL 2024, https://aclanthology.org/2024.acl-long.123/)

2. **Can LLMs Approximate Human Judgment Distributions Using Explanations?** (EMNLP 2024, https://aclanthology.org/2024.findings-emnlp.842/)

3. **Can LLMs Replace Humans in Generating Explanations?** (ACL 2025, https://aclanthology.org/2025.findings-acl.562/

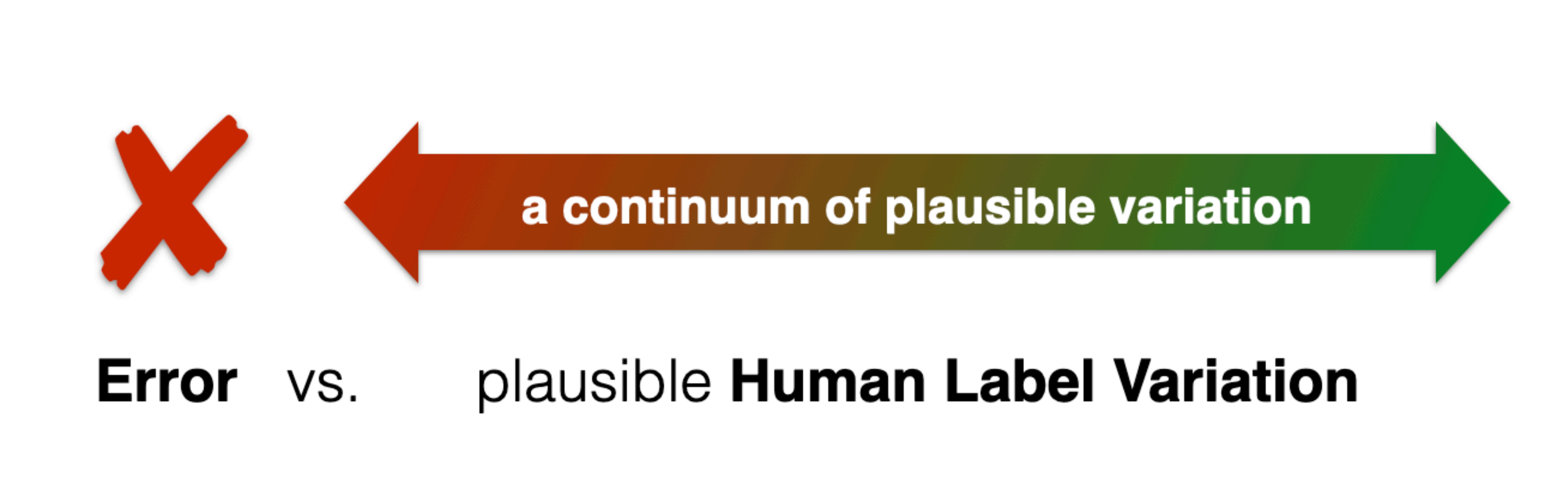
4. **Can We Use Linguistic Taxonomy to Better Understand Explanations?** (EMNLP 2025, https://arxiv.org/abs/2505.22848)

RQ1:
How to Separate Annotation Error from Human Label Variation?

VariErr NLI: Separating Annotation Error from Human Label Variation

Leon Weber-Genzel[▲]* **Siyao Peng**[▲]* **Marie-Catherine de Marneffe**[✒] **Barbara Plank**[▲]
[▲] MaiNLP & MCML, LMU Munich, Germany
[✒] FNRS, CENTAL, UCLouvain, Belgium
{siyao.peng,b.plank}@lmu.de marie-catherine.demarneffe@uclouvain.be

While Human Label Variation (HLV) exists, so do errors.



Can we tease apart annotation error from plausible human label variation?

VariErr: Variation vs. Error

- We provide one theoretical and operational answer: **VariErr.**
- Two rounds of annotation:
- **Round 1: annotators provide free-text explanations for their labels during annotation;**
- **Round 2: all annotators independently judge R1 label-explanation pairs;**
- This mirrors traditional annotation adjudication but without agreeing on a single label/explanation.

Label	Annotator	Explanation
Entailment	1	the picture hit Tommy in the head
	2	a picture hit Tommy's head with terrific force
	3	Tommy hurt his head with the picture
Neutral	3	ambiguous if Tommy hurt himself or another guy
Contradiction	4	Tommy is not hurt but rather bad strong emotion

L	A	Expl.	*Validity Judgment* 1	2	3	4
E		…				
	3	Tommy hurt …	✓	✓	✓	✗
N	3	ambiguous …	✓	✓	✓	✗
C	4	Tommy is …	✗	✗	✓	✗

VariErr Round 2: Validating Labels & Explanations

- **Self judgment:** judging his/her **own** Round 1 annotations;
- **Peer judgment:** judgments from **the others annotators;**
- Conventionally, annotations that **were originally different but changed to an agreeing label** after adjudication are **corrected errors**;
- We define **annotation error** as: **annotator him/herself invalidate his/her R1 annotations.**

L	A	Expl.	*Validity Judgment*				*Self-validated?*	*Error?*
			1	2	3	4		
E	1	`the picture …`	✓	✓	✓	✗	Yes, ✓	
	2	`a picture …`	✓	✓	✓	✗	Yes, ✓	*Not an error, 3* ✓
	3	`Tommy hurt …`	✓	✓	✓	✗	Yes, ✓	
N	3	`ambiguous …`	✓	✓	✓	✗	Yes, ✓	*Not an error, 1* ✓
C	4	`Tommy is …`	✗	✗	✓	✗	No, ✗	*An error, 0* ✓

Automatic Error Detection (AED) on VariErr

- Backgrounds: most AEDs rely on **post-hoc error mining or injecting synthetic errors**; VariErr provides **gold-labelled errors.**
- **Task: determine whether an NLI label is an error** given (premise, hypothesis);
- **Models**: AED models and LLMs; as well as human heuristics — Label Count (LC) and Peer judgments.
- Observation 1: **peer vote** is the best estimate;
- Observation 2: **GPT4** is the best model (given additional access to **explanation** annotations).

Scorer	AP	P@100	R@100
	Baselines		
Random	14.7	14.7	11.4
	Models		
MA	17.7	18.3	14.2
DM_{mean}	22.8	23.7	18.3
DM_{std}	22.3	22.7	17.6
GPT-3.5	17.6	21.0	16.3
GPT-4	**31.3**	**46.0**	**35.9**
	Human		
LC_{CHAOS}	32.5	35.0	27.3
LC_{VARIERR}	40.8	42.0	32.6
$Peer_{avg}$	42.2	46.0	35.9
$Peer_{sum}$	**46.5**	**47.0**	**36.7**

We found it crucial
to investigate explanations
in NLI annotations!

RQ2:
Can LLMs Approximate Human Judgment Distributions Using Explanations?

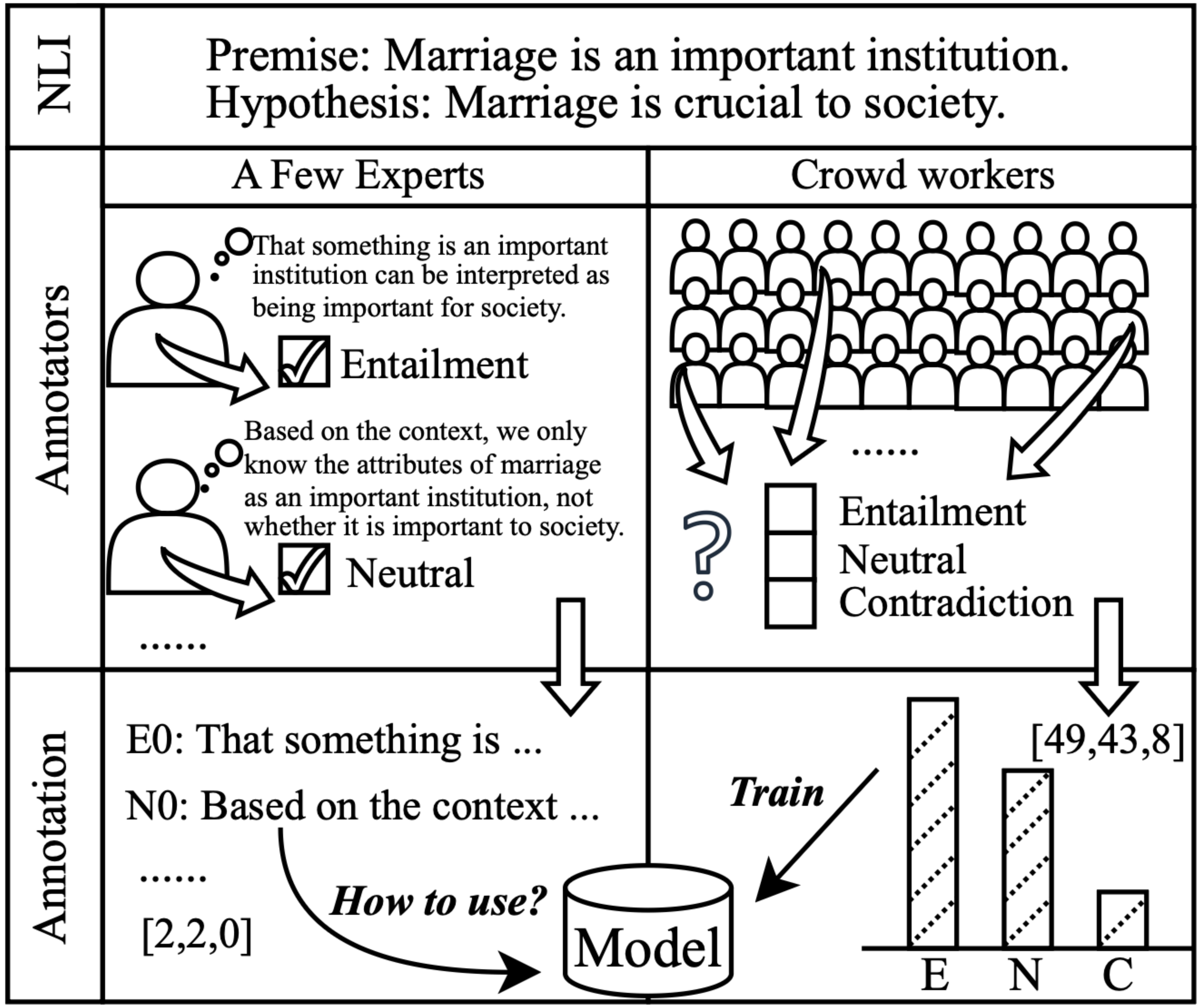
"Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations?

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

MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
Language Technology Lab, University of Cambridge, United Kingdom
{beiduo.chen, xinpeng.wang, siyao.peng, robert.litschko}@lmu.de
alk23@cam.ac.uk, b.plank@lmu.de

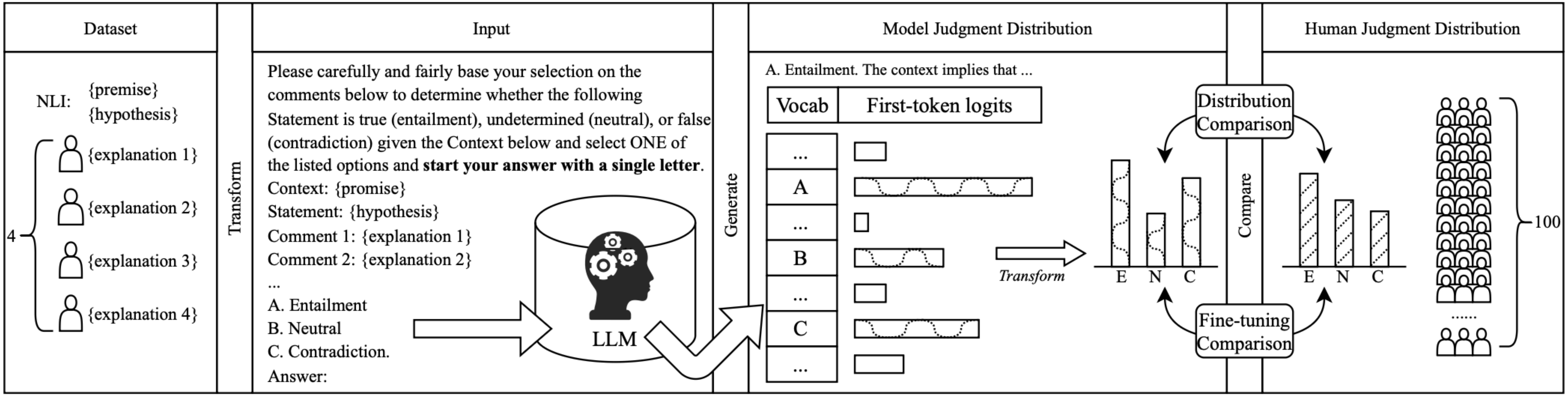
A Few Explanations vs. Many Labels?

- Many NLI datasets exhibit HLV:
    - ChaosNLI: 100 crowd annotations per item
    - VariErrNLI: 4 annotators with explanations for each label
- **Can we approximate Human Judgement Distribution** (HJD, from ChaosNLI) **using a few labels and explanations** (from VariErrNLI)?



Distribution and Fine-tuning Comparisons

- **Distribution Comparison:** How well LLM-derived Model Judgment Distributions (MJDs) approximate Human Judgment Distributions (HJDs)?
- **Fine-tuning Comparison:** How well the resulting MJDs approximate human labels when fine-tuning smaller language models (BERT/RoBERTa)?



Prompt Scenarios

- Without explanations
- With explanations
- With explicit explanations (incl. labels)

Type	General Instruction Prompt
Without explanations	**"role": "user", "content"**: Please 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} A. Entailment B. Neutral C. Contradiction. Answer:
With explanations	**"role": "user", "content"**: 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:

With explicit explanations	**"role": "user", "content"**: 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}, so I choose {label 1} Comment 2: {explanation 2}, so I choose {label 2} ... A. Entailment B. Neutral C. Contradiction. Answer:

Distribution Comparison

- We compare LLM-derived MJDs to the HJDs from ChaosNLI;
- Measures: Kullback-Leibler (KL) Divergence, Jensen Shannon Distance (JSD), and Total Variation Distance (TVD);
- Observation: **a few explanations can improve the capabilities of LLMs** to approximate HJD.

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**

Fine-tuning Comparison

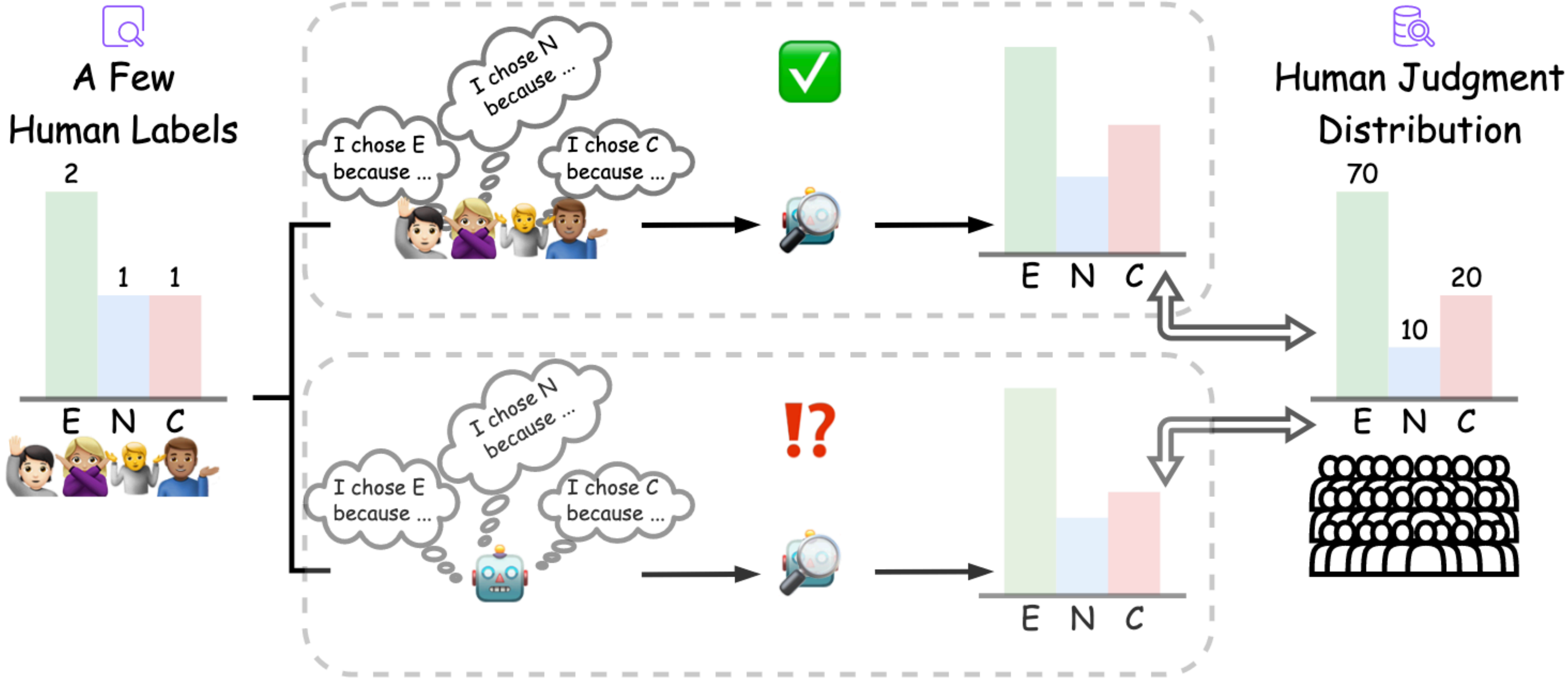
- Measures: KL and weighted F1;
- Observation: **adding explicit explanations contributes to the best models.**

Distributions	BERT FT (dev / test)		
	Weighted F1 ↑	KL ↓	CE Loss ↓
Baseline			
Chaos NLI train set	**0.626 / 0.646**	**0.074 / 0.077**	**0.972 / 0.974**
MNLI single label	0.561 / 0.589	0.665 / 0.704	2.743 / 2.855
MNLI distributions	0.546 / 0.543	0.099 / 0.102	1.046 / 1.048
VariErr distributions	0.557 / 0.559	0.179 / 0.186	1.286 / 1.299
MJDs from Mixtral			
p_{norm} of Mixtral	0.416 / 0.422	0.134 / 0.133	1.152 / 1.142
+ "serial" explanations	0.443 / 0.454	0.145 / 0.141	1.183 / 1.166
+ "serial" explicit explanations	0.506 / 0.511	0.130 /0.130	1.139 / 1.132
+ "parallel" explanations	0.404 / 0.428	0.134 / 0.131	1.150 / 1.136
+ "parallel" explicit explanations	**0.507 / 0.514**	**0.108 / 0.108**	**1.074 / 1.065**
p_{sfmax} of Mixtral	0.427 / 0.432	0.131 / 0.129	1.140 / 1.130
+ "serial" explanations	0.452 / 0.462	0.121 / 0.118	1.113 / 1.096
+ "serial" explicit explanations	0.509 / **0.520**	0.105 / 0.105	1.064 / 1.057
+ "parallel" explanations	0.397 / 0.429	0.121 / 0.119	1.112 / 1.098
+ "parallel" explicit explanations	**0.522** / 0.517	**0.095 / 0.095**	**1.035 / 1.026**

RQ3:
Can LLMs Replace Humans in Generating Explanations?

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

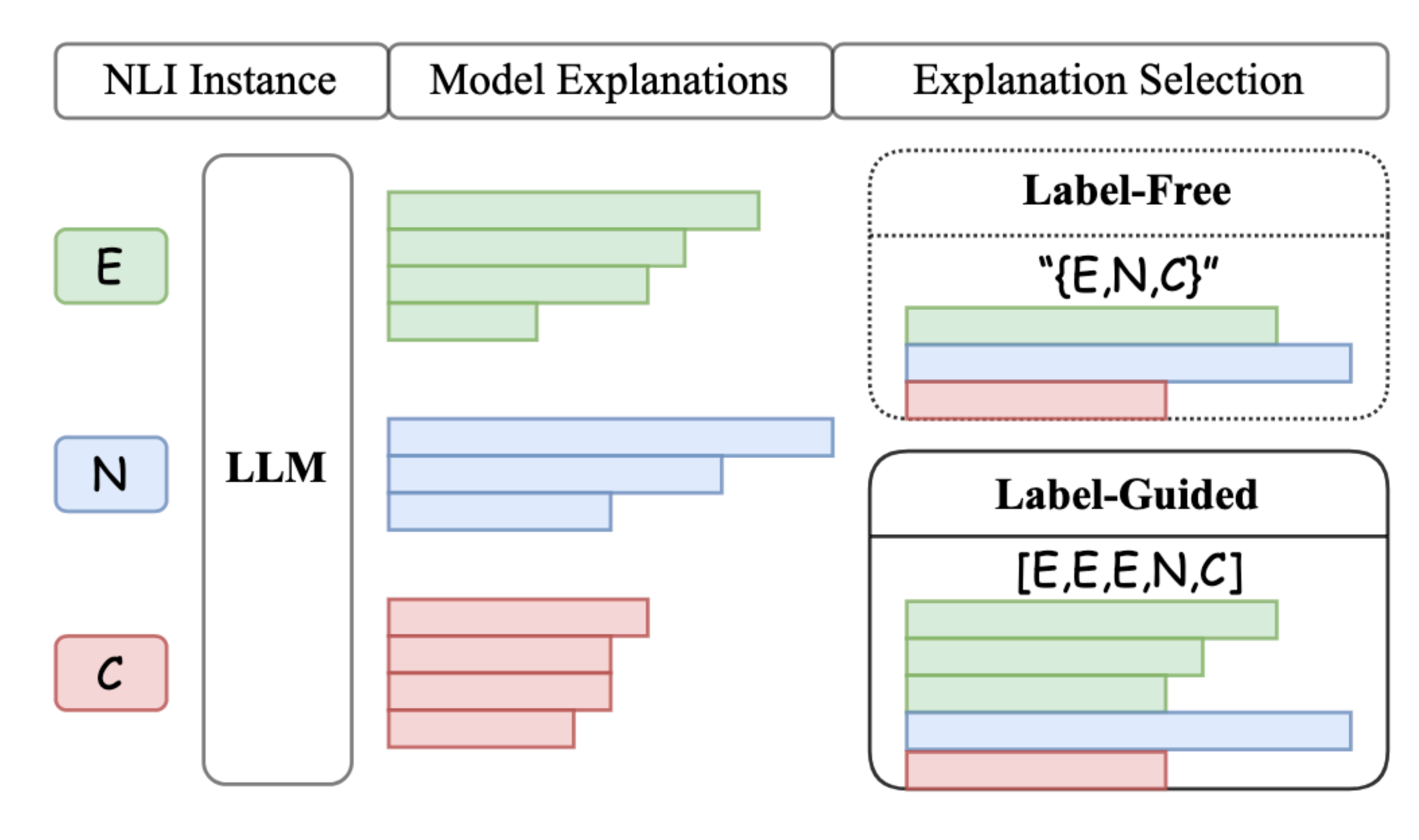
Beiduo Chen **Siyao Peng** **Anna Korhonen** **Barbara Plank**
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
Language Technology Lab, University of Cambridge, United Kingdom
beiduo.chen@lmu.de, siyao.peng@lmu.de, alk23@cam.ac.uk, b.plank@lmu.de



A Few
Human Labels
2
1
1
E N C
I chose E because ...
I chose N because ...
I chose C because ...
E N C
I chose E because ...
I chose N because ...
I chose C because ...
E N C
Human Judgment
Distribution
70
10
20
E N C

Explanation Selection

- We prompt LLMs to **generate as many explanations as possible for each label**;
- We experiment with two explanation selection strategies:
- **Label-Free:** using one explanation for each of the three NLI labels;
- **Label-Guided:** selecting explanations based on the annotated NLI labels.



Results

- Ceiling: ChaosNLI HJD;
- **Label-Free:** minor improvements;
- **VariErr Label-Guided model explanation achieves comparable results** to LLMs with human explanations;
- **“A rose by any other name would smell as sweet”** — William Shakespeare's play Romeo and Juliet.

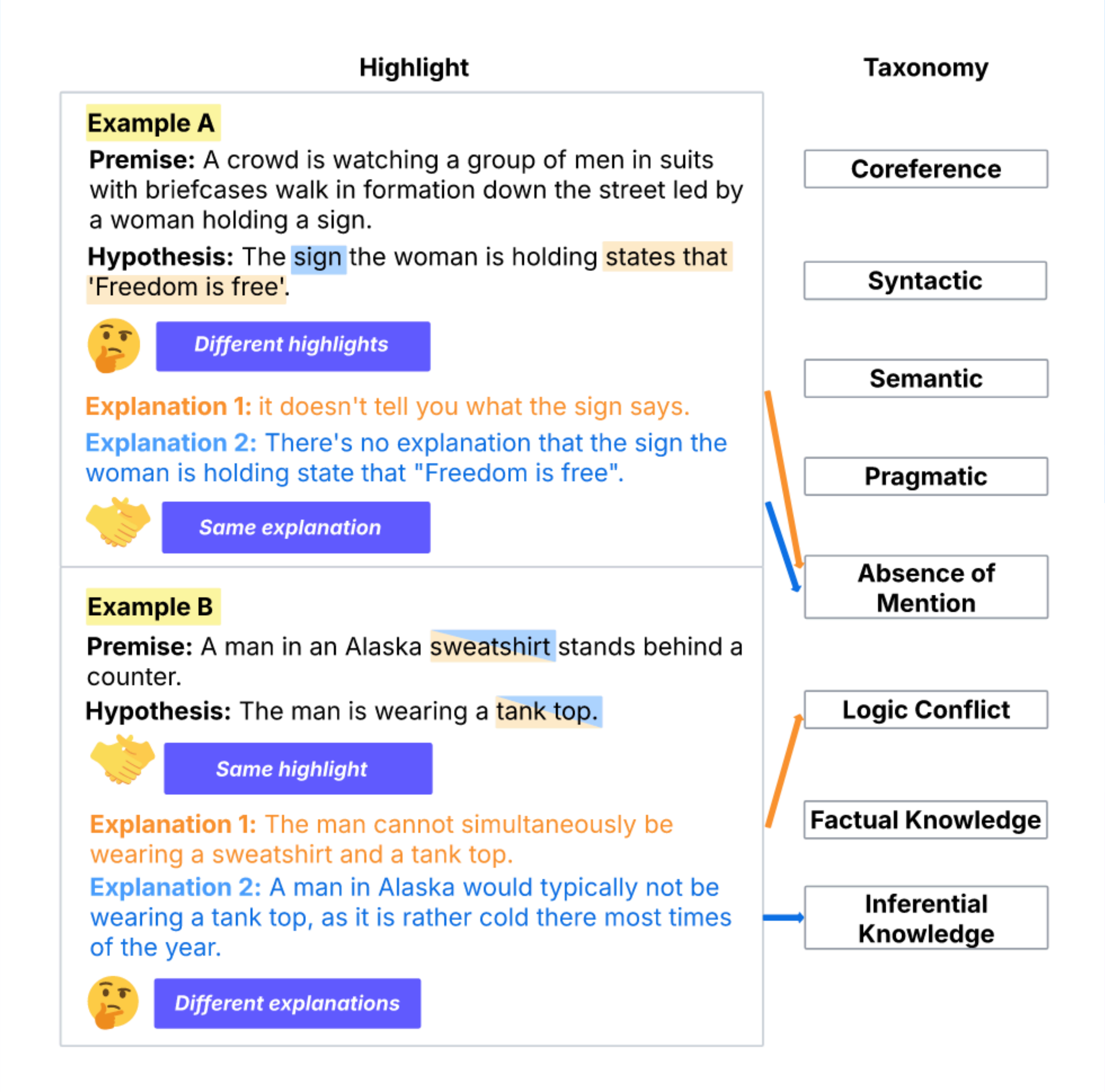
Distributions	Dist. Comparison			BERT Fine-Tuning Comparison (dev/test)		
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑
Baseline from Human Annotations						
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609
VariErr distribution	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522
MNLI distribution	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555
Model Judgment Distributions						
Llama3	0.259	0.262	0.284	0.099 / 0.101	1.045 / 1.044	0.516 / 0.487
+ human explanations	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.039	0.575 / 0.556
+ model explanations						
Label-Free	0.295	0.278	0.310	0.106 / 0.107	1.066 / 1.063	0.539 / 0.533
VariErr Label-Guided	**0.234**	**0.247**	**0.266**	0.097 / 0.098	1.041 / 1.037	0.558 / 0.544
MNLI Label-Guided	0.242	0.251	0.275	**0.096 / 0.097**	**1.037 / 1.034**	**0.589 / 0.580**
GPT-4o	0.265	0.263	0.289	0.103 / 0.096	1.059 / 1.029	0.526 / 0.517
+ human explanations	**0.187**	**0.207**	**0.223**	0.093 / 0.098	1.027 / 1.036	**0.570 / 0.552**
+ model explanations						
Label-Free	0.252	0.242	0.275	0.101 / 0.102	1.052 / 1.047	0.537 / 0.545
VariErr Label-Guided	0.192	0.209	0.226	**0.092 / 0.093**	**1.026 / 1.022**	0.554 / 0.551

We have shown that human/LLM explanations help capture HLV.

But how do we evaluate the similarities among explanations?

Similar Explanations?

- Explanations are **free texts;**
- Our earlier papers used **lexical, syntactic and semantic** measures;
- **Token highlighting** were used as proxies;
- **But none of these signal that two explanation sentences essentially mean the same thing.**



RQ4:
Can We Use Linguistic Taxonomy to Better Understand Explanations?

LITEx: A LInguistic Taxonomy of Explanations for Understanding Within-Label Variation in Natural Language Inference

Pingjun Hong[*†] **Beiduo Chen**[*] **Siyao Peng** **Marie-Catherine de Marneffe** **Barbara Plank**

MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning, Germany FNRS, CENTAL, UCLouvain, Belgium
Faculty of Computer Science and UniVie Doctoral School Computer Science, University of Vienna, Austria
pingjun.hong@univie.ac.at, {beiduo.chen, siyao.peng, b.plank}@lmu.de, marie-catherine.demarneffe@uclouvain.be

LiTeX

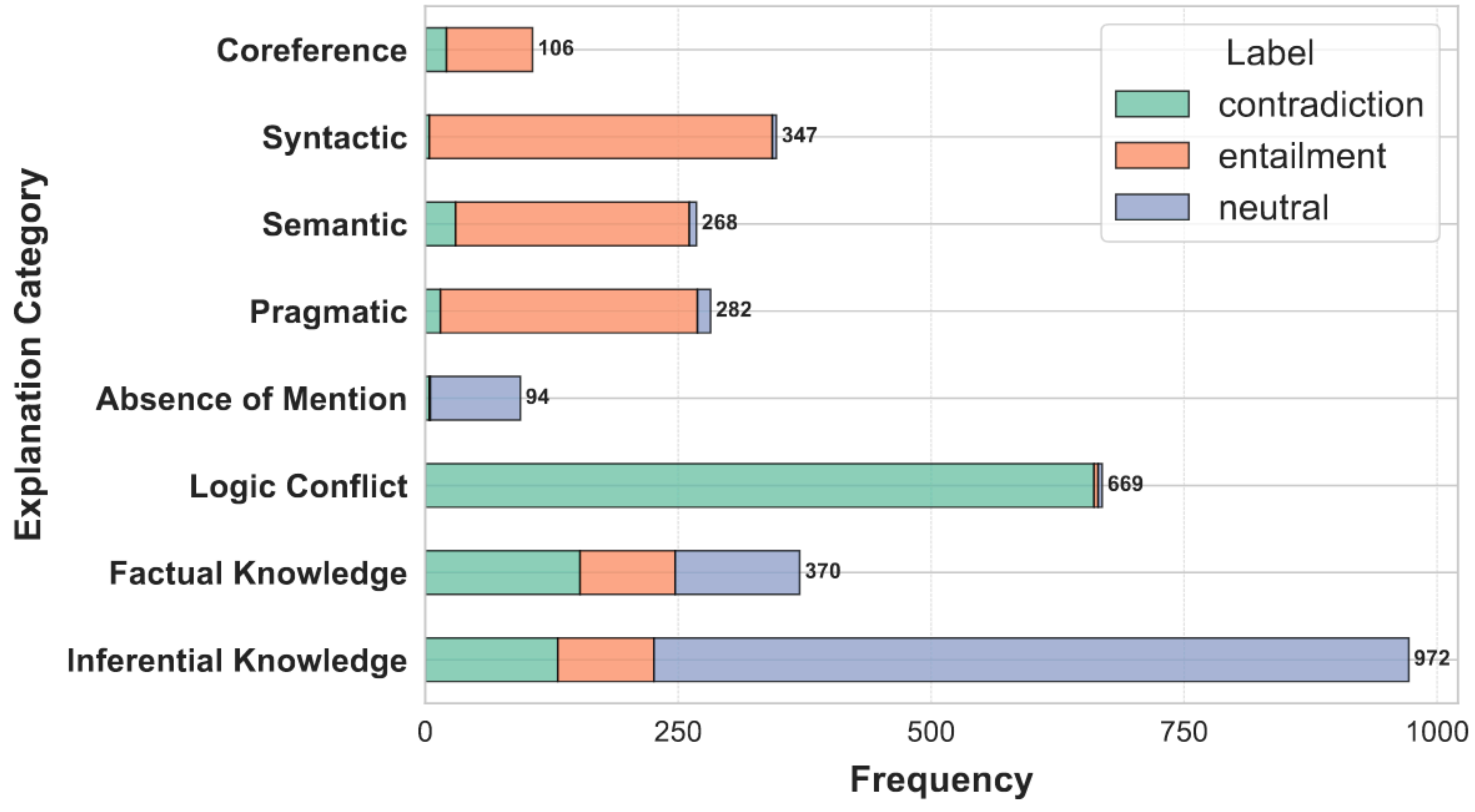
- LiTeX: A **Li**nguistic **T**axonomy of **eX**planations
- Two broad categories:
- **Text-Based (TB) Reasoning:** explanations depending solely on **surface-level linguistic evidence** found within (P, H);
- **World-Knowledge (WK) Reasoning:** explanations that invoke **background knowledge or domain-specific information** beyond text.

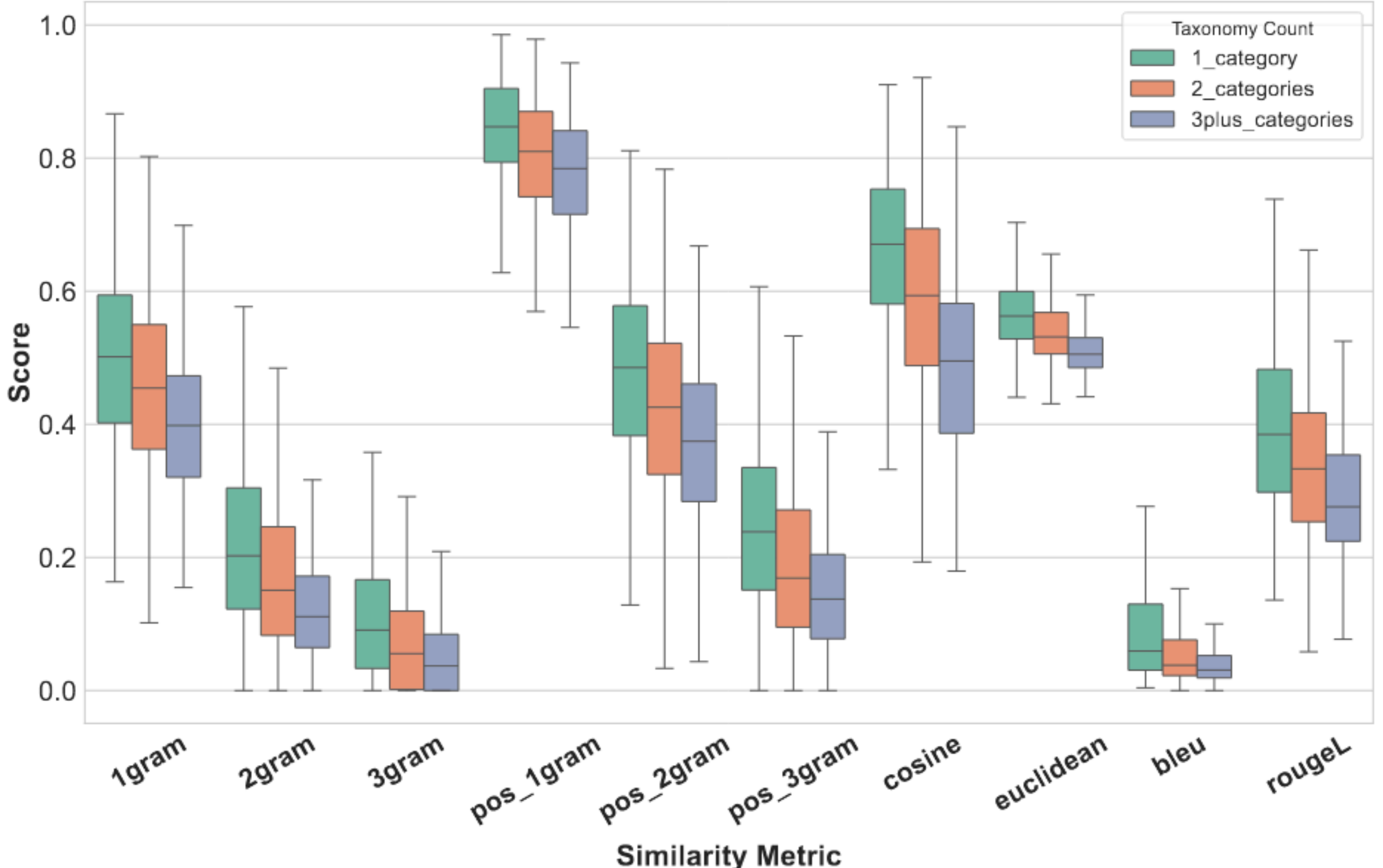
		Text-Based Reasoning (TB)
Coreference	**Q:** **Check:**	*Does the explanation rely on resolving coreference between entities?* Determine whether the main entities in the premise and hypothesis refer to the same real-world referent, including via pronouns or phrases.
Syntactic	**Q:** **Check:**	*Does the explanation involve a change in sentence structure that preserves meaning?* Determine whether the premise and hypothesis differ in structure, such as active vs. passive, reordered arguments, or coordination/subordination, while preserving the same meaning.
Semantic	**Q:** **Check:**	*Does the explanation involve semantic similarity or substitution of key concepts?* Evaluate whether core words or expressions - including verbs, nouns, and adjectives - are semantically related between the premise and hypothesis. This includes synonymy, antonymy, lexical entailment, or category membership.
Pragmatic	**Q:** **Check:**	*Does the explanation rely on pragmatic cues like implicature or presupposition?* Look for meaning beyond the literal text - including implicature, presupposition, speaker intention, and conventional conversational meaning.
Absence of Mention	**Q:** **Check:**	*Does the explanation point out information not mentioned in the premise?* Check whether the hypothesis introduced information that is neither supported nor contradicted by the premise - i.e., it is not mentioned explicitly.
Logic Conflict	**Q:** **Check:**	*Does the explanation refer to logical constraints or conflict?* Evaluate whether the hypothesis interacts with the premise via logical structures, such as exclusivity, quantifiers ("only", "none"), or conditionals, which constrain or conflict with each other.
		World Knowledge-Based Reasoning (WK)
Factual Knowledge	**Q:** **Check:**	*Does the explanation rely on widely shared, intuitive facts acquired through everyday experience?* Determine whether the explanation invokes commonly known facts, such as physical properties or universal experiences, that are not stated in the premise.
Inferential Knowledge	**Q:** **Check:**	*Does the explanation rely on real-world norms, customs, or culturally grounded reasoning?* Determine whether the explanation requires reasoning based on general world knowledge, including cultural expectations, social norms, or typical causal inferences, that are not stated in the premise.

Table 1: Guiding questions and decision criteria for our LITEX taxonomy.

Taxonomy Analysis

- **Connecting NLI labels to LiTeX Categories:**
    - Logic Conflict ~ contradiction
    - Syntactic/Semantic/Pragmatic ~ entailment
    - Absence of Mention ~ neutral
    - Factual/Inferential Knowledge more even
- **Explanation similarity decreases as the number of different taxonomy categories on an NLI item increases.**





Generating Explanations Using LiTeX

- **Goal:** to generate multiple explanations **that reflect different plausible reasoning paths** for a given NLI item and its label.
- **Prompt designs:**
  - **Baseline:** only premise, hypothesis and label;
  - **Highlight-Guided:** additionally **highlight annotations** on (P, H);
  - **Taxonomy-Guided:** additionally **taxonomy description**, one example for each of the eight reasoning categories from LiTeX.

Results

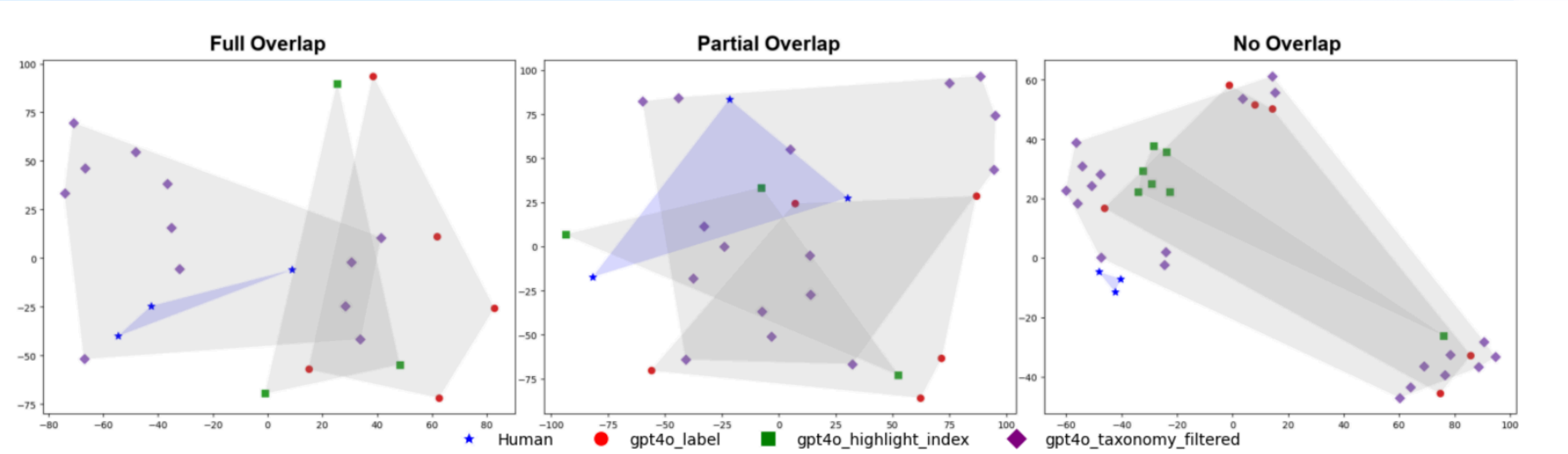
- **Taxonomy prompting performs best** on three LLMs;
- **Highlight-guided generations tend to be verbose** (longer explanations) and yielding **lower BLEU and ROUGE-L scores**.

Mode		Word n-gram			POS n-gram			Semantic		NLG Eval		Avg_len
		1-gram	2-gram	3-gram	1-gram	2-gram	3-gram	Cos.	Euc.	BLEU	ROUGE-L	
GPT4o	*baseline*	0.291	0.117	0.049	0.882	0.488	0.226	0.556	0.524	0.051	0.272	24.995
	highlight (indexed)	0.402	0.124	0.053	0.878	0.481	0.222	0.554	0.522	0.051	0.269	28.240
	taxonomy (two-stage)	0.418	0.128	0.071	0.886	0.495	0.242	0.593	0.537	0.071	0.314	**19.991**
	taxonomy (end-to-end)	**0.437**	**0.166**	**0.083**	**0.898**	**0.511**	**0.255**	**0.608**	**0.540**	**0.074**	**0.323**	26.672
DeepSeek-v3	*baseline*	0.369	0.087	0.034	0.847	0.449	0.195	0.428	0.490	0.042	0.245	**20.288**
	highlight (indexed)	0.364	0.091	0.037	0.861	0.450	0.196	0.464	0.499	0.034	0.242	27.301
	taxonomy (two stage)	0.391	0.122	0.055	0.884	0.475	0.219	0.544	0.522	0.057	0.293	20.894
	taxonomy (end-to-end)	**0.404**	**0.140**	**0.067**	**0.897**	**0.486**	**0.233**	**0.556**	**0.528**	**0.063**	**0.306**	25.960
Llama-3.3-70B	*baseline*	0.392	0.106	0.044	0.863	0.478	0.224	0.466	0.496	0.046	0.250	27.148
	highlight (indexed)	0.317	0.065	0.024	0.807	0.408	0.173	0.367	0.478	0.031	0.199	24.987
	taxonomy (two-stage)	**0.444**	**0.167**	**0.082**	0.889	**0.512**	**0.256**	**0.609**	**0.541**	**0.078**	**0.321**	**22.340**
	taxonomy (end-to-end)	0.383	0.110	0.048	**0.896**	0.499	0.232	0.505	0.510	0.047	0.262	28.870

Table 4: Similarity of LLM-generated explanations to human references.

How much variation can LLM-generated explanations cover?

- Are LLMs **too repetitive** and only cover **a subset** of human explanations?
- Can LLMs **unearth appropriate new explanations** that are missing from a few human-written ones?
- **full coverage:** the t-SNE convex hull of model-generated explanations **fully encloses all human explanation points**; (2) partial coverage, and (3) no coverage.



Measures

- **Full Coverage:** if **all** human explanation reference points **fall within** the convex hull spanned by the model explanations;
- **Partial Coverage:** if **at least one** human reference point is within the model explanation space;
- **Area Precision: the ratio of the overlapping area** over the area spanned by **model explanations**;
- **Area Recall:** the ratio of the overlapping area over the area spanned by **human explanations**.

Results

Mode	Coverage Full	Coverage Partial	Area Rec	Area Prec
GPT4o *baseline*	1.9	21.6	16.5	**5.7**
highlight (indexed)	1.1	13.5	10.0	4.7
taxonomy (end-to-end)	**10.7**	**56.1**	**49.3**	5.6
DeepSeek-v3 *baseline*	4.0	20.5	17.5	2.7
highlight (indexed)	2.3	14.9	12.5	2.9
taxonomy (end-to-end)	**17.8**	**61.8**	**54.7**	**3.8**
Llama-3.3-70B *baseline*	1.7	15.4	12.2	2.9
highlight (indexed)	0.5	8.2	6.5	2.5
taxonomy (end-to-end)	**16.7**	**65.2**	**59.8**	**5.7**

Taxonomy-guided explanation generation consistently achieves **the highest coverage** of reference explanation points**,** as well as the highest **average area recall and precision.**

HLV —> Explanations —> What's next?

Ongoing Work

- Deconstructing label variation in NLI through **explanations**
- **Individual Variability** in NLI
- Can **LLMs valid their explanations and labels**? Do LLMs represent opinions from **a single person** or **a group of people**?
- How does **multilinguality and cultural variation** affect label variation?
- What is a **good** explanation — prominent entities, conciseness, casual relations, etc.?
- Still an open area …

Thank you

Backup — Discussion & Future

- **Did GPT-4 perform well due to ChaoNLI in training?**
- Not really! GPT-4 AED does not solely rely on ChaosNLI — mid Pearson r correlation.

- **Can we combine Label Counts with Training Dynamics?**
- Yes! Via reranking, we observe that combining HLV with AEDs is promising.

We found it crucial to investigate explanations in NLI annotations!