"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

Outline

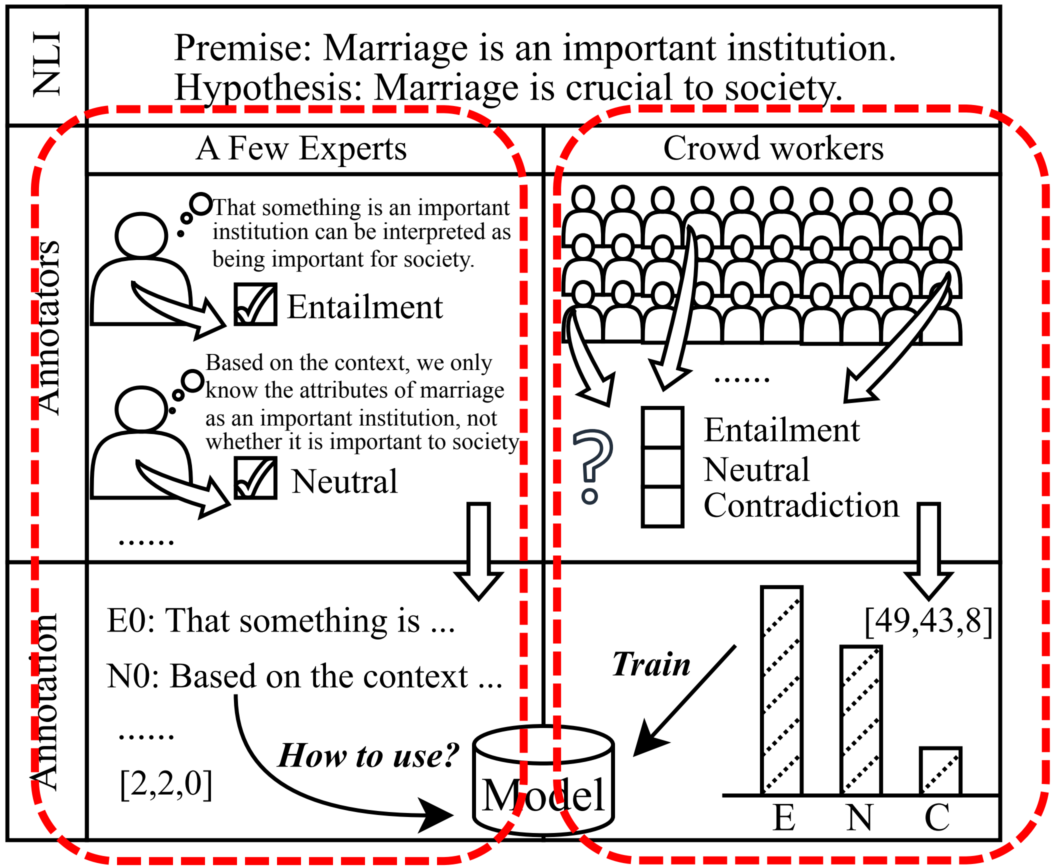
- **Introduction**
- LLMs to Estimate Human Judgment Distributions
- Experimental Setup
- Results & Discussion
- Conclusion

Introduction

- **Human Label Variation** (HLV) is a valuable source of information that arises when multiple human annotators provide different labels for valid reasons.

- In **Natural Language Inference**

Introduction



Introduction

- **Human Label Variation** (HLV) is a valuable source of information that arises when multiple human annotators provide different labels for valid reasons.

- In **Natural Language Inference**, approaches to capturing HLV involve either collecting annotations from many crowd workers to human judgment distribution (HJD) or use expert linguists to provide detailed explanations for their chosen labels.

- **Large Language Models** (LLMs) are increasingly used as evaluators ("LLM judges") but with mixed results, and few works aim to study HJDs.

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

LLMs to Estimate HJDs

- Research Question:

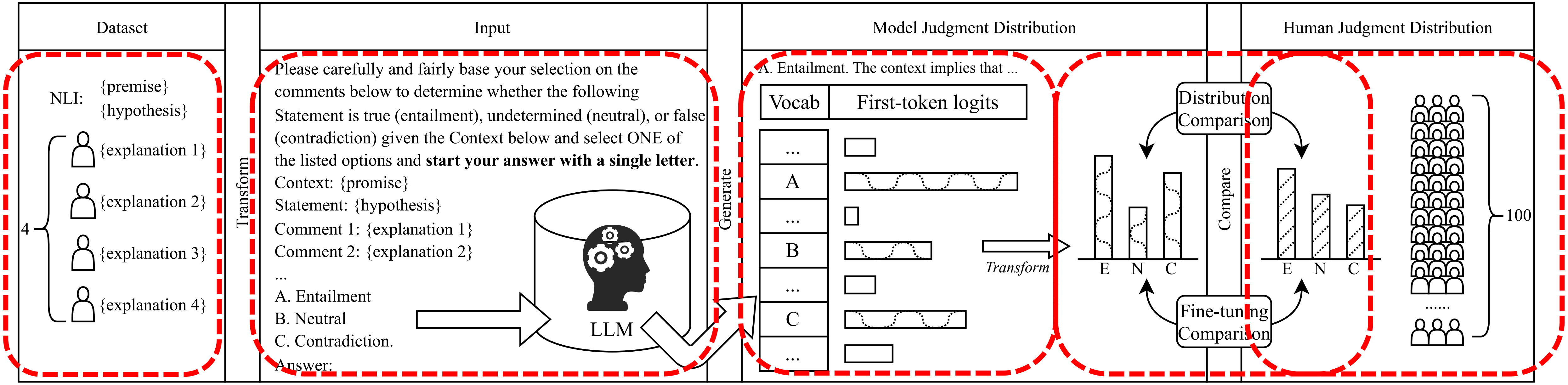
- *1. Can LLMs provided with a "small" number of detailed explanations better approximate the human judgment distributions collected by a "big" number of annotators?*

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

Outline

- Introduction
- **LLMs to Estimate Human Judgment Distributions**
- Experimental Setup
- Results & Discussion
- Conclusion

LLMs to Estimate HJDs



LLMs to Estimate HJDs

- First-token Probability

$$p^{O}_{\text{norm}}(j) = \frac{s_j}{\sum_{j}^{|O|} s_j},$$

$$p^{O}_{\text{sfmax}}(j) = \frac{\exp(s_j/\tau)}{\sum_{j}^{|O|} \exp(s_j/\tau)},$$

- Bias Consideration
  - ABC orders; explanation orders
  - Serial / Parallel processing mode

- With/Without Explicit Label

Outline

- Introduction
- LLMs to Estimate Human Judgment Distributions
- **Experimental Setup**
- Results & Discussion
- Conclusion

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Experimental Setup

- Distribution Comparison
  - MJDs vs. HJDs
    - Kullback-Leibler Divergence (KL)
    - JensenShannon Distance (JSD)
    - Total Variation Distance (TVD)

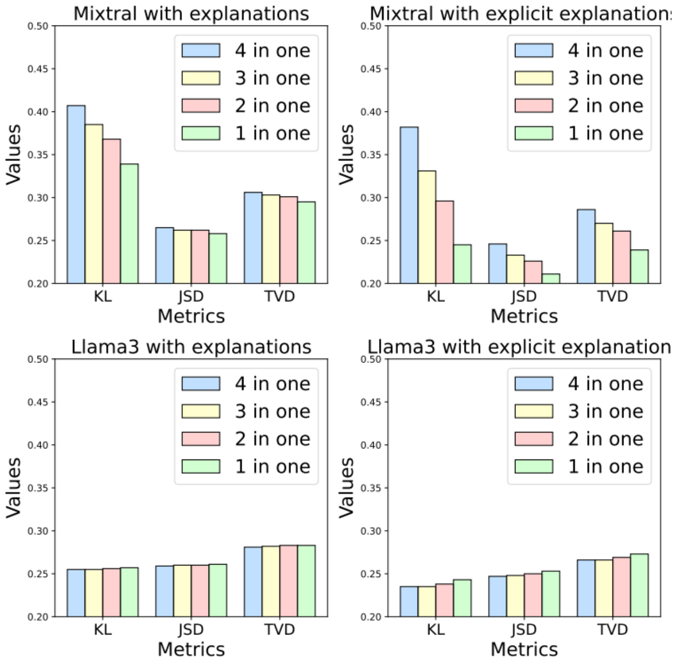
- Fine-tuning Comparison
  - Soft-label fine-tuning on BERT/RoBERTa : MJDs vs HJDs
    - Kullback-Leibler Divergence (KL)
    - Cross-Entropy Loss (CE Loss)
    - Weighted F1 score

Outline

- Introduction
- LLMs to Estimate Human Judgment Distributions
- Experimental Setup
- **Results & Discussion**
- Conclusion

Results & Discussion

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



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

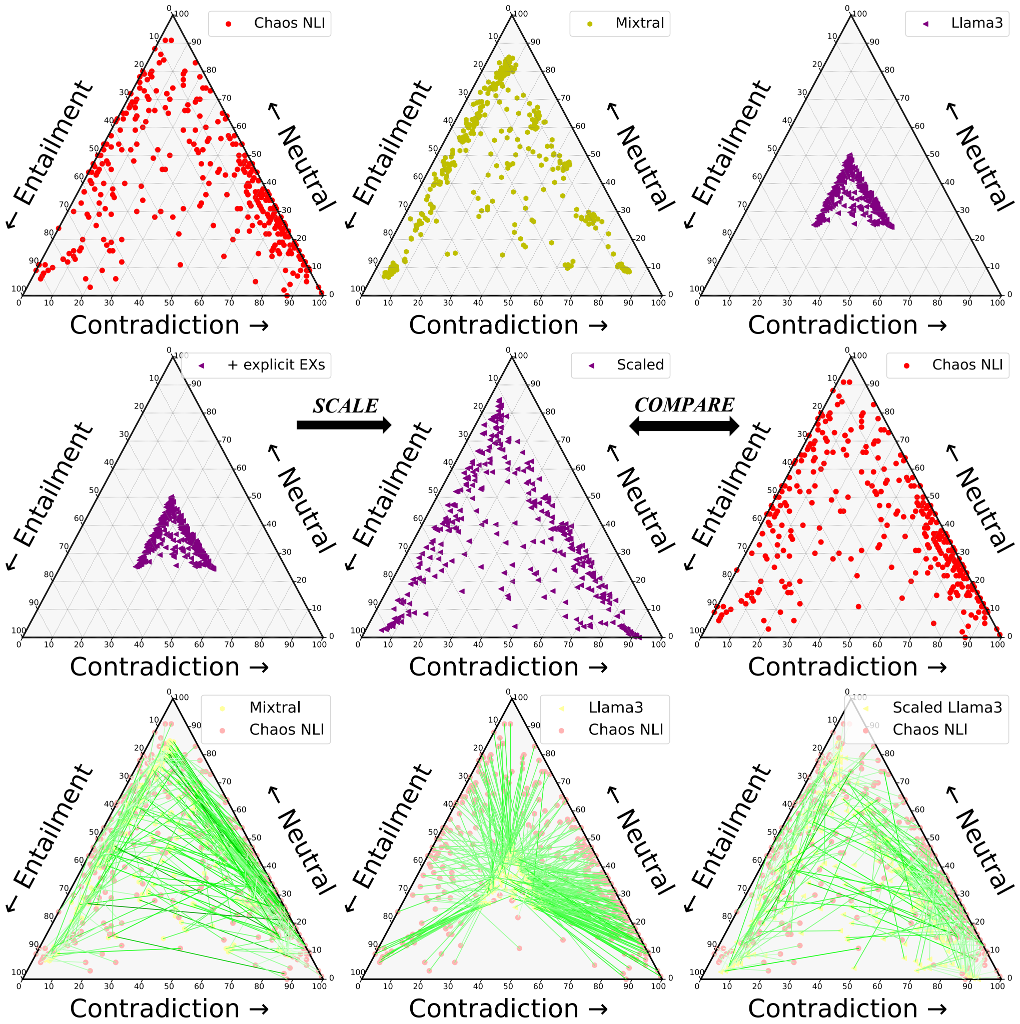
Results & Discussion

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 & Discussion



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**

Outline

- Introduction
- LLMs to Estimate Human Judgment Distributions
- Experimental Setup
- Results & Discussion
- **Conclusion**

Conclusion

- Explanation works.
- FT Comparison *cannot* be predicted well by Dist. Comparison.
- Llama3 and Mixtral exhibit rather different clusters in visualization. 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 aligned with the HJD than Mixtral and supports its better fine-tuning performances.
- 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.

Thank you !!!

Presenter: Beiduo Chen
Email: Beiduo.Chen@lmu.de

Resource:

Paper

Code

Acknowledgement:

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